Compressive Sensing Using Symmetric Alpha-Stable Distributions – Part II: Robust Sparse Reconstruction

George Tzagkarakis*, John P. Nolan, and Panagiotis Tsakalides, *Member, IEEE*

***Abstract*—Traditional compressive sensing (CS) primarily assumes light-tailed models for the underlying signal and/or noise statistics. Nevertheless, this assumption is not met in the case of highly impulsive environments, where non-Gaussian infinite variance processes arise for the signal and/or noise components. This drives the traditional sparse reconstruction methods to failure, since they are incapable of suppressing the effects of heavy-tailed sampling noise. In the companion paper (Part I), we proposed a robust nonlinear sampling operator, which mitigates the effects of impulsive observation noise by employing a generalized alpha-stable matched filter for the generation of random measurements. The family of symmetric alpha-stable (SαS) distributions, as a powerful tool for modeling heavy-tailed behaviors, is also adopted in this paper to design a robust sparse reconstruction algorithm from noisy random measurements. Specifically, a novel greedy reconstruction method is developed, which achieves increased robustness to impulsive sampling noise by solving a minimum dispersion (MD) optimization problem based on fractional lower-order moments. The MD criterion emerges naturally in the case of additive sampling noise modeled by SαS distributions, as an effective measure of the spread of reconstruction errors around zero, due to the lack of second-order moments. The experimental evaluation demonstrates the improved reconstruction performance of the proposed algorithm when compared against state-of-the-art CS techniques for a broad range of impulsive environments.**

***Index Terms*—Compressive sensing, sparse recovery, symmetric alpha-stable distributions, heavy-tailed statistics, fractional lower-order moments, minimum dispersion criterion.**

I. Introduction

USING the concept of transform coding, compressive sensing (CS) enables a potentially large reduction in the sampling and computation costs for capturing signals that have a *sparse* or *compressible* representation. Furthermore, CS is characterized by an intrinsic denoising mechanism, which suppresses the non-sparse contributions due to noise. This is extremely important in practical applications, where an observed signal and/or its corresponding set of compressive measurements are typically corrupted by noise.

This work was partially funded by EONOS Investment Technologies, and the European Union's Horizon 2020 Research and Innovation Programme PHySIS under grant agreement No. 640174.

G. Tzagkarakis is with EONOS Investment Technologies, Paris, France, e-mail: gtzag@eonos.com.

J. P. Nolan is with the Math/Stat Department, American University, Washington, DC, e-mail: jpnolan@american.edu. Supported by contract W911NF-12-1-0385 from the U.S. Army Research Office.

P. Tsakalides is with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS) and the Department of Computer Science, University of Crete, Heraklion, Greece, e-mail: tsakalid@ics.forth.gr.

In order to suppress the noise impact, which deteriorates the accurate reconstruction of the original signal from a reduced set of compressive measurements, a broad range of noise-aware sparse reconstruction algorithms have been developed. These include: greedy pursuit [1], [2], convex relaxation [3], [4], [5], Bayesian formulation [6], [7], nonconvex optimization [8], [9], [10] and brute force [11]. Each method has its own advantages and limitations. For instance, greedy pursuits and convex optimization are computationally more tractable and yield provably correct reconstructions under well-determined conditions. However, apart from sparsity, they are not able to account for any prior statistical information about the signal and noise, which could be used to improve the reconstruction accuracy. On the other hand, Bayesian methods and nonconvex optimization are typically based on rigorous principles, but often they do not provide theoretical guarantees. Finally, brute force approaches are algorithmically solid, but their practical use is restricted to small-scale problems.

The majority of previous CS reconstruction methods is primarily based on light-tailed, finite-variance assumptions for the underlying statistics of the signal and/or noise generating processes. Despite the analytical tractability and practical appeal, these assumptions may yield a dramatic degradation of the reconstruction quality when we operate in highly impulsive environments, which give rise to *heavy-tailed* processes with infinite variance. To alleviate the effects of gross errors that mask the information conveyed by the compressive measurements, recent state-of-the-art methods rely on algebraic-tailed models, in particular, on the Cauchy and generalized Cauchy (GCD) distributions, to design robust reconstruction methods [12], [13].

On the other hand, *alpha-stable* distributions [14] have been proven very powerful in accurately modeling impulsive phenomena. However, their intractability due to the lack of closed-form expressions for the density functions of all except for a few stable distributions (Gaussian, Cauchy and Lévy) has prevented their exploitation in the framework of compressive sensing. To address this problem, whilst also revealing the advantages of alpha-stable models in designing efficient CS systems, this paper and its companion paper (Part I) [15] propose an integrated framework for robust nonlinear sampling under heavy-tailed observation noise (Part I) and robust sparse signal reconstruction under heavy-tailed sampling noise, by modeling the noise statistics via *symmetric alpha-stable* (SαS) distributions. To the best of our knowledge, this is the first thorough study that bridges the fields of CS and SαS models for the design of a complete CS system. We emphasize that the methodologies proposed herein (for sparse signal reconstruction) and the companion paper (for compressive sampling) can be used either independently of each other, or together to form an integrated CS system with increased robustness to heavy-tailed infinite-variance noise.

A. Motivation

In practical CS acquisition systems, the generated compressive measurements are typically corrupted by *sampling* noise. The presence of large-amplitude noise in the measurement domain degrades dramatically the reconstruction accuracy of traditional CS techniques based on ℓ_1 or ℓ_2 norms. The problem becomes even more challenging when we operate in impulsive environments, where the corrupting sampling noise can be of infinite variance. This causes conventional sparse reconstruction algorithms to fail in recovering a close approximation of the original signal.

The problem of accurate sparse signal reconstruction from random measurements corrupted by gross sampling errors has been previously addressed in the context of error correction coding [16] and incomplete measurements [17], [18]. The main limitation of these approaches is that their reconstruction performance relies on the sparsity of the error term, which is a condition that may not be met often in practice.

Recently, the problem of robust sparse signal reconstruction from random measurements corrupted by impulsive sampling noise has been addressed efficiently by employing the Lorentzian norm either as an objective function or as a constraint. In particular, in [12], [19] the true sparse signal is reconstructed by solving an ℓ_0-regularized least logarithmic deviation problem. The logarithmic deviation is defined in terms of the Lorentzian norm, which does not over-penalize large deviations, and is therefore more robust than the commonly used ℓ_1 and ℓ_2 norms for the suppression of impulsive noise. A similar approach is proposed in [13], where a nonconvex optimization problem is solved to reconstruct the sparse signal by minimizing the ℓ_1 norm subject to a nonlinear constraint based on the Lorentzian norm. The use of the Lorentzian norm in these papers is further justified by the existence of logarithmic moments for heavy-tailed distributions, since the second-order moments are infinite or even undefined for such distributions.

Despite the enhanced reconstruction quality of the above methods in the presence of heavy-tailed additive sampling noise, the specific use of Cauchy [12], [19] or GCD distributions [13] can be restrictive in capturing more generic non-Gaussian heavy-tailed behaviors of the sampling noise. Motivated by this limitation, first we model the statistics of highly impulsive sampling noise, with possibly infinite variance, by members of the SαS family. Then, we propose a novel greedy sparse reconstruction algorithm, which minimizes the *dispersion* of the random measurements' error. As it will become clear in the subsequent analysis, the minimum dispersion (MD) criterion arises naturally as a measure of the spread of estimation errors around zero for random variables modeled by SαS distributions, as is the case with the infinite variance sampling noise adopted in this study. Most importantly, we show that the MD criterion is equivalent to a minimum ℓ_p estimation error criterion, which simplifies the design of our proposed sparse reconstruction algorithm. Nevertheless, we emphasize that, without loss of generality, the subsequent analysis considers a conventional linear sampling operator for the generation of random measurements. However, our proposed reconstruction algorithm is generic and can be combined with any sampling operator (e.g. the nonlinear operator proposed in the companion paper [15]).

B. Main Contributions

The major contribution of this paper is twofold: i) we propose a novel iterative greedy algorithm that combines the characteristics of a gradient-descent approach with a statistical optimization criterion, namely, the *minimum dispersion* (MD) criterion, which is equivalent to minimizing the *fractional lower-order moments* (FLOMs) of reconstruction errors. The FLOMs measure the ℓ_p ($p < 2$) distance between the reconstructed and the true sparse signal; ii) we provide theoretical guarantees for the convergence of the algorithm and an upper bound of the ℓ_2 norm of the reconstruction error, along with rules of thumb for setting the key parameters that control the performance of the proposed algorithm.

Furthermore, the use of ℓ_p distance metrics with $p < 2$, which arise naturally when SαS models are coupled with a minimum dispersion rule, provides an additional degree of freedom (i.e., the value of p) yielding increased robustness against gross sampling errors. The proposed reconstruction algorithm resembles an orthogonal matching pursuit (OMP) approach in the sense that, at each step, it selects the measurement basis vector which is most correlated with the current residuals. However, the key difference between our algorithm and an OMP-based approach is that this correlation is expressed in terms of FLOMs, thus it adapts perfectly to non-Gaussian, heavy-tailed processes with infinite variance. At each iteration, one or several elements of the sparse signal are reconstructed, therefore, as the algorithm progresses, a refined estimate of its nonzero elements is obtained by removing the contribution of previously estimated elements.

C. Paper Organization

The rest of the paper is organized as follows: Section II introduces the minimum dispersion as a proper optimization criterion for heavy-tailed, infinite variance sampling noise modeled by SαS distributions, and shows its equivalence with a minimum ℓ_p estimation error criterion. Section III analyzes the design and implementation of our proposed iterative greedy algorithm for sparse signal reconstruction. Furthermore, it provides theoretical proofs for the convergence of the algorithm, along with an upper bound of the ℓ_2 norm of the reconstruction error. An experimental evaluation of the reconstruction accuracy of our proposed algorithm is presented in Section IV for a variety of impulsive environments, where its performance is compared against state-of-the-art sparse reconstruction methods tailored to impulsive sampling noise. Finally, Section V concludes and gives directions for future work.

D. Notation

In the following, scalars are denoted by lower-case letters (e.g. x), column vectors by lower-case boldface letters (e.g. $\mathbf{x}$), and matrices by upper-case boldface letters (e.g. $\mathbf{X}$). Sets are represented by calligraphic letters (e.g. $\mathcal{S}$), while $|\mathcal{S}|$ denotes their cardinality. The ith column of a matrix $\mathbf{X}$ is denoted by $\mathbf{x}_i$, whereas x_j indicates the jth element of a vector $\mathbf{x}$. $\mathcal{S}_i$ denotes a subset of set $\mathcal{S}$, while $\mathbf{X}_\mathcal{S}$ designates the submatrix formed by the columns $\{\mathbf{x}_i \mid i \in \mathcal{S}\}$, whose indices belong to $\mathcal{S}$. Similarly, $\mathbf{x}_\mathcal{S}$ denotes the subvector formed by the elements $\{x_j \mid j \in \mathcal{S}\}$, whose indices belong to $\mathcal{S}$. Finally, we use $\hat{\mathbf{x}}$, $\mathbf{x}^\mathsf{T}$, $\mathbf{x}^*$, and $\mathbf{x}^{(t)}$ to denote the estimate (reconstruction), transpose, optimal solution, and value at tth iteration of a vector $\mathbf{x}$, respectively. Similar notations are used for the matrices.

II. Sparse Reconstruction and Minimum Dispersion Criterion

Let $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\mathsf{T} \in \mathbb{R}^N$ be a real discrete-time signal. In the general case, we assume that $\mathbf{x}$ can be sparsified over a (possibly overcomplete) transform basis $\mathbf{\Psi} \in \mathbb{R}^{N' \times N}$ with $N' \geq N$, such that $\boldsymbol{\alpha} = \mathbf{\Psi}\mathbf{x} \in \mathbb{R}^{N'}$ is an s-sparse vector of transform coefficients. $\mathbf{\Psi}$ and $\mathbf{\Psi}^\mathsf{T}$ denote the analysis (direct) and synthesis (inverse) transforms, respectively.

For convenience, a linear sampling operator is employed for the generation of a reduced set of compressive measurements. Nevertheless, we emphasize again that our proposed reconstruction method can be equally combined with nonlinear sampling operators, such as those proposed in [12], [13] or in the companion paper [15].

Given a random measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ ($M < N$), which satisfies all the necessary and sufficient conditions for accurate sparse reconstruction, the generic noisy sampling model adopted in the subsequent analysis is as follows,

$$\mathbf{y} = \mathbf{\Phi}\mathbf{\Psi}^\mathsf{T}\boldsymbol{\alpha}_0 + \mathbf{n} \ , \tag{1}$$

where $\boldsymbol{\alpha}_0 \in \mathbb{R}^{N'}$ is the true sparse signal to be recovered, $\mathbf{y} \in \mathbb{R}^M$ is a vector of M noisy random measurements, and $\mathbf{n} \in \mathbb{R}^M$ is the additive sampling noise. By setting $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}^\mathsf{T}$, various well-established approaches for recovering the sparse signal solve a constrained optimization problem of the form,

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{N'}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}\|_p \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_q \leq s \ , \tag{2}$$

or a regularized optimization problem,

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{N'}} \left(\|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}\|_p + \tau\|\boldsymbol{\alpha}\|_q\right) \ , \tag{3}$$

with $0 < q \leq 1$, $p \in \{2, \infty\}$ or denoting the Lorentzian norm[1], s being the sparsity level and τ a regularization parameter that balances the influence of the data fidelity term and the sparsity-inducing term on the optimal solution. Having obtained an estimate of the optimal sparse coefficients vector $\boldsymbol{\alpha}_0^*$, the original signal is given by $\hat{\mathbf{x}}_0 = \mathbf{\Psi}^\mathsf{T}\boldsymbol{\alpha}_0^*$.

[1] The ℓ_p norm of a vector is defined by $\|\mathbf{x}\|_p = \left(\sum_{j=1}^N |x_j|^p\right)^{1/p}$ for $0 < p \leq 2$ (ℓ_p is a quasi-norm for $0 < p < 1$), whilst the Lorentzian norm is defined by $\|\mathbf{x}\|_{LL_2} = \sum_{j=1}^N \log(1 + x_j^2)$.

At the core of our proposed sparse reconstruction algorithm is the use of SαS distributions for modeling the statistics of impulsive sampling noise. In particular, we assume that the additive noise $\mathbf{n} \sim f_{\alpha_n}(\gamma_n, 0)$ is a random variable that follows a univariate SαS distribution with characteristic exponent $\alpha_n \in (0, 2]$, dispersion $\gamma_n > 0$ and zero location parameter (ref. Section II-B in the companion paper [15]). Moreover, the smaller the α_n, the heavier the tails of the noise density function.

Due to their algebraic tails, SαS distributions lack finite second-order moments. Instead, all moments of order p less than α do exist and are called the *fractional lower-order moments* (FLOMs). In particular, the FLOMs of a SαS random variable $X \sim f_\alpha(\gamma_X, 0)$ are given by [14]

$$\mathbb{E}\left\{|X|^p\right\} = \left(C_{p,\alpha} \cdot \gamma_X\right)^p \ , \quad 0 < p < \alpha \ , \tag{4}$$

where

$$(C_{p,\alpha})^p = \frac{2^{p+1}\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(-\frac{p}{\alpha}\right)}{\alpha\sqrt{\pi}\,\Gamma\left(-\frac{p}{2}\right)} \left(= \frac{\Gamma\left(1 - \frac{p}{\alpha}\right)}{\cos\left(\frac{\pi}{2}p\right)\Gamma(1 - p)}\right) \ . \tag{5}$$

In addition, (4) yields the following expression for the dispersion of X in terms of the FLOMs,

$$\gamma_X = \left(\mathbb{E}\left\{|X|^p\right\}\right)^{1/p} C_{p,\alpha}^{-1} \ , \tag{6}$$

which will be employed to quantify the spread of gross noise samples around zero.

Furthermore, due to the lack of finite variance for alpha-stable distributed data, the traditional minimum mean squared error (MMSE) criterion cannot be used as a measure of the reconstruction quality to be optimized. Instead, we employ the *minimum dispersion* (MD) criterion in our optimization problem, since the dispersion of alpha-stable random variables is finite and gives a good measure of the spread of estimation errors around zero. Most importantly, our proposed reconstruction method belongs to the class of ℓ_p-based nonlinear, non-convex relaxation techniques. Specifically, ℓ_p (quasi-)norms ($0 < p < 2$), and subsequently the use of a *least ℓ_p estimation error* criterion, emerge naturally in the case of infinite variance sampling noise modeled as a SαS random variable. Indeed, as we deduce from (6), an ℓ_p (quasi-)norm based approximation of the dispersion of X can be obtained by replacing the FLOM, $\mathbb{E}\{|X|^p\}$, with a discrete finite sum,

$$\mathbb{E}\{|X|^p\} \approx \frac{1}{N}\sum_{j=1}^N |x_j|^p = \frac{1}{N}\|\mathbf{x}\|_p^p \ . \tag{7}$$

Let $\hat{X}$ denote an estimate of the random variable X and $E = X - \hat{X}$ be the estimation error. Then, the minimum dispersion criterion can be viewed as a minimum ℓ_p estimation error criterion. Clearly, from (6) and (7) we deduce that minimizing the dispersion of the error, γ_E, is equivalent to minimizing the ℓ_p (quasi-)norm of the associated error vector $\mathbf{e} \in \mathbb{R}^N$, which is considered as a realization of E.

1) Estimation of p: Notice that the selection of an appropriate value for the p parameter in the above expressions is a critical step. Most importantly, the optimal p depends on the noise characteristic exponent, α_n, which is estimated from the noisy compressive measurements. A method for choosing the optimal p as a function of α has been proposed in [20], which is based on minimizing the standard deviation of a FLOM-based covariation estimator[2]. Other authors suggest the optimal p should be lower but as close as possible to the value of α, if α can be inferred. However, the later approach entails the divergence of the pth FLOM as $p \to \alpha$, since $C_{p,\alpha}$ in (4) goes to infinity. On the other hand, the former approach yields an almost linear relation between α and the optimal value of p, and specifically $p \lesssim \alpha/2$. In addition, if $p < \alpha/2$ the FLOM estimator has a finite variance, which is desirable [21]. In the following, the optimal value of p is set as a function of α by linearly interpolating the entries of the lookup Table I, generated by minimizing the standard deviation of a FLOM-based covariation estimator.

TABLE I
OPTIMAL p PARAMETER AS A FUNCTION OF THE CHARACTERISTIC EXPONENT α.

α	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
p_{opt}	0.52	0.56	0.58	0.61	0.64	0.69	0.72	0.76	0.81	0.88	0.98

III. Robust Sparse Reconstruction Algorithm

In this section, we propose a novel iterative greedy reconstruction algorithm, which suppresses efficiently the effects of heavy-tailed, infinite variance sampling noise, whilst achieving increased robustness to a broader range of impulsive noise behaviors, from linear (i.e., Gaussian) to extremely impulsive sampling noise. First, we demonstrate the inefficiency of traditional sparse reconstruction methods to deal with impulsive sampling noise. Then, we analyze the design and implementation of our proposed algorithm, which solves an ℓ_0-constrained ℓ_p^p optimization problem.

For convenience, yet without loss of generality, we assume that the original signal $\mathbf{x}_0$ is sparse in the canonical basis $\mathbf{\Psi} = \mathbf{I}$, thus $\mathbf{x}_0 = \boldsymbol{\alpha}_0$. It will be mentioned explicitly whenever the original signal is assumed to be sparse in a different transform basis, that is, when $\mathbf{\Psi} \neq \mathbf{I}$.

A. *Effects of Heavy-Tailed Sampling Noise on Reconstruction Accuracy*

When traditional sparse reconstruction methods are applied on linear random measurements, the *oracle estimator* achieves the best reconstruction performance, without any prior knowledge about the distribution of the original sparse signal. Nevertheless, the oracle estimator has perfect knowledge of the support of $\mathbf{x}_0$, that is, the set $\mathcal{T} \subset \{1, 2, \ldots, N\}$ of indices of the s nonzero elements of $\mathbf{x}_0$. Then, by assuming power-limited sampling noise, an estimate of the sparse signal is obtained by taking the least-squares projection onto the subspace spanned by the columns of $\mathbf{\Phi}$ (or $\mathbf{A}$ in the generic case when $\mathbf{\Psi} \neq \mathbf{I}$) with indices in $\mathcal{T}$.

[2]Covariances do not exist in the family of SαS random variables due to their infinite variance. Instead, a quantity called *covariation*, which plays an analogous role for SαS random variables to the one played by covariance for Gaussian random variables has been proposed [14].

Let $\mathbf{\Phi}$ be a valid random measurement matrix with rows of norm $\sqrt{\lambda}$, which satisfies the restricted isometry property (RIP) of order s with constant (RIC) δ_s, and $\sigma_{\mathbf{n}}^2$ denote the variance of the sampling noise. When the measurements are corrupted by Gaussian sampling noise, the following inequality holds for the MSE of the oracle estimator [22],

$$\mathbb{E}\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2\right\} \geq \frac{s \cdot \sigma_{\mathbf{n}}^2}{1+\delta_s}\,, \tag{8}$$

where $\mathbb{E}\{\cdot\}$ denotes the expected value of a random variable.

From (8) it becomes obvious that traditional sparse reconstruction methods are inefficient when the random measurements are corrupted by heavy-tailed noise. Indeed, the expected reconstruction error for the oracle estimator can be very large when $\sigma_{\mathbf{n}}^2$ is large, or even infinite, as is the case of heavy-tailed infinite-variance noise.

Fig. 1 illustrates the incapability of traditional CS reconstruction methods to address the case of impulsive sampling noise. To this end, we consider a sparse vector $\boldsymbol{\alpha}_0$ of length $N = 1024$ with $s = 21$ randomly chosen nonzero elements drawn from the standard Cauchy distribution, as shown in Fig. 1a. The original signal, shown in Fig. 1b, is given by $\mathbf{x}_0 = \mathbf{\Psi}^{\mathsf{T}}\boldsymbol{\alpha}_0$, where $\mathbf{\Psi}^{\mathsf{T}}$ is the $N \times N$ discrete cosine transform (DCT) matrix. In Fig. 1c, a set of $M = 256$ linear measurements, $\mathbf{y}_0 = \mathbf{\Phi}\mathbf{x}_0$, is shown for the noiseless case, with $\mathbf{\Phi} \in \mathbb{R}^{256\times1024}$ whose entries are i.i.d. standard Gaussian samples and its columns are normalized to unit ℓ_2 norm. Then, a single outlier of amplitude $\eta = 10$ is added to a randomly chosen measurement, as shown in Fig. 1d. Figs. 1e-1f show the corresponding sparse vectors $\boldsymbol{\alpha}_0$ reconstructed from $\mathbf{y}_0$ and $\mathbf{y}$, respectively, using the NESTA algorithm[3], which is a fast and robust first-order method that solves Basis Pursuit (BP) problems. Clearly, NESTA achieves a perfect reconstruction performance given the noiseless measurements, yielding a signal-to-error ratio (SER) of $178.24\,\text{dB}$. However, the algorithm fails when the measurements are corrupted by impulsive sampling noise, resulting in a SER of $2.81\,\text{dB}$. The SER (in dB) is defined by

$$\mathrm{SER}(\mathbf{x}, \hat{\mathbf{x}}) = 10\log_{10}\left(\frac{\sum_{j=1}^{N} x_j^2}{\sum_{j=1}^{N}(x_j - \hat{x}_j)^2}\right). \tag{9}$$

B. *ℓ_0-Constrained Dispersion Minimization*

In the following, we aim at designing a robust sparse reconstruction algorithm, expressed as an operator $\mathrm{R} : \mathbb{R}^M \longmapsto \mathbb{R}^N$, which reconstructs accurately the true sparse signal $\mathbf{x}_0$ from a highly reduced set of noisy linear measurements $\mathbf{y}$. This algorithm must be robust, especially when we operate in highly impulsive environments, in the sense that small perturbations to the noiseless measurements should yield small perturbations in the reconstructed signal, even when a portion of the measurements is corrupted by large-amplitude noise.

Our proposed approach relies on the use of a least ℓ_p estimation error criterion ($0 < p < 2$), which is equivalent to a minimum dispersion criterion (ref. Section II) for SαS-distributed sampling noise. Furthermore, the presence of the

[3]NESTA toolbox: http://statweb.stanford.edu/~candes/nesta/nesta.html.

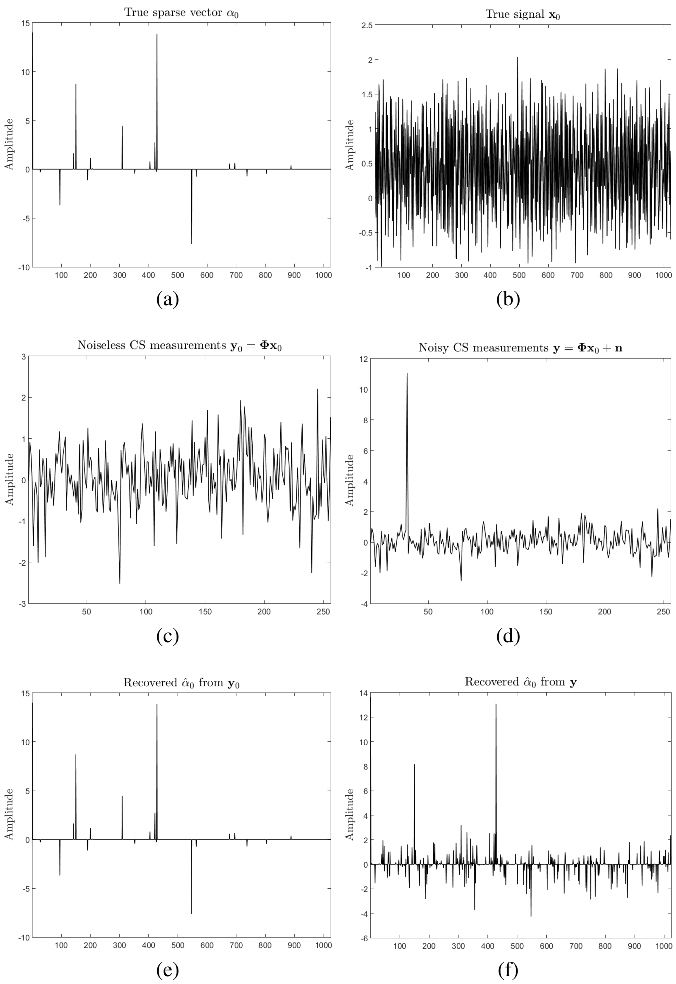


Fig. 1. Reconstruction of an impulsive signal when the random linear measurements are corrupted by a single outlier. NESTA algorithm is used to solve a BP problem. (a) True sparse vector; (b) original DCT synthesized signal; (c) noiseless linear measurements; (d) noisy linear measurements corrupted by a single outlier; (e) reconstructed sparse vector from noiseless measurements; (f) reconstructed sparse vector from noisy measurements.

free parameter p in the ℓ_p-based penalization of the residual increases the robustness against gross outliers, as opposed to the recently used Lorentzian norm. This is because, in contrast to the Lorentzian that is intrinsically related with a Cauchy model (fixed $\alpha = 1$), in our method the value of p depends on the inherent impulsiveness of the noise as expressed by its estimated characteristic exponent.

More specifically, the original sparse signal $\mathbf{x}_0$ is recovered by minimizing the dispersion of the data fidelity term, constrained on the maximum number of nonzero elements of $\mathbf{x}_0$, that is,

$$\min_{\mathbf{x}\in\mathbb{R}^N} \gamma^p_{\mathbf{y}-\mathbf{\Phi x}} \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq s \ , \tag{10}$$

where $\gamma_{\mathbf{y}-\mathbf{\Phi x}}$ is the dispersion of the data fidelity term, which is simply the difference between the original and the reproduced measurements.

From (6) and (7), the above optimization problem is expressed in the following equivalent form,

$$\min_{\mathbf{x}\in\mathbb{R}^N} \left(\frac{C^{-p}_{p,\alpha}}{M} \|\mathbf{y} - \mathbf{\Phi x}\|^p_p \right) \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq s \ . \tag{11}$$

By noticing that for the feasible range of α ($0 < \alpha \leq 2$) and p ($0 < p < \alpha$) the constant $\frac{1}{M} C^{-p}_{p,\alpha}$ is always positive (see Appendix A), the original sparse signal $\mathbf{x}_0$ is reconstructed by solving an ℓ_0-constrained least ℓ^p_p optimization problem,

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{\Phi x}\|^p_p \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq s \ . \tag{12}$$

The ℓ^p_p-based formulation arises naturally from the inherent infinite variance statistics of the impulsive sampling noise, while it also adapts effectively to less demanding light-tailed environments. Nevertheless, the numerical solution of the optimization problem in (12) can be extremely complex, even for moderate signal dimensions, due to the presence of the ℓ_0 constraint. Therefore, the design of a computationally tractable, yet very efficient, algorithm is imperative to solve the above nonconvex, combinatorial, sparse recovery problem.

C. Gradient Projection Iterative Hard Thresholding

In this section, we derive a suboptimal approach to solve (12) by employing a gradient projection (GP) formulation combined with an iterative hard thresholding (IHT) algorithm. This approach does not require the computationally intense process of matrix inversion, while providing near-optimal error guarantees [23].

More specifically, let $\mathbf{x}^{(t)}$ denote the estimated sparse solution at tth iteration, and set the initial solution to the zero vector, $\mathbf{x}^{(0)} = \mathbf{0}$. Furthermore, we are given the measurements vector $\mathbf{y}$ and the measurement matrix $\mathbf{\Phi}$. Note that in the general case, where the original signal is not sparse by itself but can be sparsified in an appropriate transform domain, the measurement matrix is replaced by $\mathbf{A} = \mathbf{\Phi \Psi}^\mathsf{T}$ and the solution corresponds to a sparse coefficients vector $\boldsymbol{\alpha}^*_0$ (ref. Section II). Then, at each iteration the algorithm computes the updated solution as follows,

$$\mathbf{x}^{(t+1)} = \mathrm{H}_s \left(\mathbf{x}^{(t)} + \mu^{(t)} \mathbf{g}^{(t)} \right) , \tag{13}$$

where $\mathrm{H}_s(\mathbf{x})$ denotes the hard thresholding operator, which sets all except for the largest (in magnitude) s elements of $\mathbf{x}$ to zero, $\mu^{(t)}$ is a step size and $\mathbf{g}^{(t)}$ is a search direction. Depending on the value of p, the ℓ^p_p-based objective function can be strictly convex ($p \geq 1$), thus the search direction can be chosen naturally as the negative gradient (gradient descent), or it can be strictly concave ($p < 1$), and the search direction can be chosen as the positive gradient (gradient ascent),

$$\mathbf{g}^{(t)} = \begin{cases} -\nabla_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|^p_p & , \text{for } p \geq 1 \\ \nabla_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|^p_p & , \text{for } p < 1 \ . \end{cases} \tag{14}$$

Notice that the hard thresholding operation may not yield a unique output. In this case, we select the s elements either at random or based on a predetermined ordering.

After some algebraic manipulation, the gradient vector is given by

$$\nabla_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|^p_p = -p\, \mathbf{\Phi}^\mathsf{T} \mathbf{W}^{(t)} \left(\mathbf{y} - \mathbf{\Phi x}^{(t)} \right) , \tag{15}$$

where $\mathbf{W}^{(t)} \in \mathbb{R}^{M \times M}$ is a diagonal matrix, whose main diagonal elements are as follows,

$$\mathbf{W}^{(t)}_{i,i} = \frac{1}{|y_i - \boldsymbol{\phi}^\mathsf{T}_i \mathbf{x}^{(t)}|^{(2-p)}} \ , \quad i = 1, \ldots, M \ . \tag{16}$$

In fact, the gradient vector can be only approximated near zero, since the inherent absolute value is not differentiable at zero. To avoid numerical instabilities in the case of a denominator close to zero, a small damping factor $\kappa > 0$ is incorporated in (16) resulting in the following weights,

$$\mathbf{W}_{i,i}^{(t)} = \frac{1}{\left((y_i - \boldsymbol{\phi}_i^{\mathsf{T}}\mathbf{x}^{(t)})^2 + \kappa\right)^{1-\frac{p}{2}}} \ , \quad i = 1, \ldots, M \ . \quad (17)$$

In the subsequent experimental evaluations we set $\kappa = 10^{-3}$. Notice also that the above weight matrix resembles the one computed when working in an iteratively reweighted least squares (IRLS) framework [24]. Furthermore, the inherent role of the weights is to suppress the effect of large errors by assigning a small weight when large deviations are estimated. In the special case of $\mathbf{W}^{(t)} = \mathbf{I}$ the above GP-based IHT method reduces to the conventional least squares IHT [23]. In the rest of the text, we will refer to our proposed algorithm for minimizing the error dispersion using a gradient projection hard thresholding approach as MD-IHT.

D. Key Parameters Setting

The efficiency of our proposed MD-IHT algorithm is affected by the accurate tuning of two key parameters, namely, the parameter p, which determines the ℓ_p^p objective function and depends on the noise characteristic exponent, and the step size $\mu^{(t)}$.

Concerning the value of p, as we have already mentioned in Section II, we adopt the almost linear relation $p \lesssim \alpha/2$. Thus the problem is reduced to estimating accurately the noise characteristic exponent from the random measurements $\mathbf{y}$. Following the approach suggested in [24], the method of log-cumulants is employed to estimate the SαS parameters by equating sample log-cumulants to their theoretical counterparts for a particular model and then solving the resulting system, much in the same way as in the classical method of moments. In particular, by applying the Mellin transform on a SαS density we get the following expression for its second-kind first characteristic function,

$$\Phi(z) = \frac{\gamma^{z-1} 2^z \Gamma\left(\frac{z}{2}\right) \Gamma\left(-\frac{z-1}{\alpha}\right)}{\alpha\sqrt{\pi}\,\Gamma\left(-\frac{z-1}{2}\right)} . \quad (18)$$

Notice that by setting $z = p + 1$ in $\Phi(z)$ we obtain the expression of the FLOMs of a SαS random variable, as defined in (4). By taking the limit as $z \to 1$ of the first and second derivatives of the $\log(\Phi(z))$, we derive the following results for the second-kind cumulants of a SαS model,

$$\begin{aligned} \tilde{k}_1 &= \frac{\alpha - 1}{\alpha}\psi(1) + \log(\gamma) \\ \tilde{k}_2 &= \frac{\pi^2}{12}\frac{\alpha^2 + 2}{\alpha^2} \ , \end{aligned} \quad (19)$$

where $\psi(\cdot)$ is the Digamma function. On the other hand, the first two sample second-kind cumulants can be estimated empirically from the M measurements $\mathbf{y}$ as follows,

$$\begin{aligned} \hat{\tilde{k}}_1 &= \frac{1}{M}\sum_{i=1}^{M} \log\left(|y_i|\right) \\ \hat{\tilde{k}}_2 &= \frac{1}{M}\sum_{i=1}^{M} \left(\log\left(|y_i|\right) - \hat{\tilde{k}}_1\right)^2 \ . \end{aligned} \quad (20)$$

The estimation process simply involves solving (19) for α and γ by substituting $\tilde{k}_1$, $\tilde{k}_2$ with their sample estimates $\hat{\tilde{k}}_1$, $\hat{\tilde{k}}_2$, respectively.

Regarding the tuning of the step size, we notice that the convergence performance of the MD-IHT algorithm improves if an adaptive step size, $\mu^{(t)}$, is employed to normalize the gradient update in (13). Specifically, let $\mathcal{S}^{(t)}$ be the support of $\mathbf{x}^{(t)}$, and assume that the algorithm has identified the true support of $\mathbf{x}_0$, that is, $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} = \mathcal{S}$. Then, we want to minimize $\|\mathbf{y} - \boldsymbol{\Phi}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_p^p$ using a gradient projection algorithm with updates

$$\mathbf{x}_{\mathcal{S}}^{(t+1)} = \mathbf{x}_{\mathcal{S}}^{(t)} + \mu^{(t)}\mathbf{g}_{\mathcal{S}}^{(t)} \ . \quad (21)$$

Optimality in estimating $\mu^{(t)}$ is equivalent to finding a step size which reduces maximally the ℓ_p^p objective function at each iteration. This is a nontrivial task and, in general, there is no known closed form for an optimal step size. To address this issue, we update the step size at each iteration in a suboptimal way as follows,

$$\begin{aligned} \mu^{(t)} &= \arg\min_{\mu} \left\| \mathbf{y} - \boldsymbol{\Phi}_{\mathcal{S}}\left(\mathbf{x}_{\mathcal{S}}^{(t)} + \mu\mathbf{g}_{\mathcal{S}}^{(t)}\right) \right\|_p^p \\ &= \arg\min_{\mu} \left\| \left(\mathbf{y} - \boldsymbol{\Phi}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{(t)}\right) - \mu\boldsymbol{\Phi}_{\mathcal{S}}\mathbf{g}_{\mathcal{S}}^{(t)} \right\|_p^p . \end{aligned} \quad (22)$$

By setting $\mathbf{u} = \left(\mathbf{y} - \boldsymbol{\Phi}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{(t)}\right)$ and $\mathbf{v} = \boldsymbol{\Phi}_{\mathcal{S}}\mathbf{g}_{\mathcal{S}}^{(t)}$, we deduce that the estimation of a suboptimal step size $\mu^{(t)}$ at the tth iteration is reduced to solving the following optimization problem,

$$\mu^{(t)} = \arg\min_{\mu} J_p(\mu) \triangleq \arg\min_{\mu} \|\mathbf{u} - \mu\,\mathbf{v}\|_p^p \ , \quad (23)$$

whose treatment depends on the value of p. Although p varies in the interval $(0, 2)$ in our proposed framework, we examine the solution of (23) in the general case of $p > 0$ by distinguishing the following cases:

i) $p = 2$: This is the simplest case, in which the optimal step size is given in closed form, corresponding to the typical least-squares (LS) minimizer, that is,

$$\mu^{(t)} = \arg\min_{\mu} J_2(\mu) = \frac{\mathbf{v}^{\mathsf{T}}\mathbf{u}}{\mathbf{v}^{\mathsf{T}}\mathbf{v}} \ . \quad (24)$$

ii) $p = 1$: In this case, the minimizer of $J_1(\mu)$ is obtained by solving a least absolute deviation (LAD) regression problem [25]. Specifically, the optimal $\mu^{(t)}$ is equal to the weighted median of the sample set $\{r_i = \frac{u_i}{v_i}\}_{i=1}^M$ with positive real weights $\{|v_i|\}_{i=1}^M$, hereafter denoted by

$$\mu^{(t)} = \mathrm{WMED}\left(|v_i| \diamond r_i \Big|_{i=1}^{M}\right) , \quad (25)$$

where $\diamond$ denotes the replication operator. In the general case of positive real weights, the weighted median can be easily computed as follows [25]:

(1) Calculate the threshold $v_\varrho = \frac{1}{2}\sum_{i=1}^{M}|v_i|$.
(2) Sort all the samples in ascending order, $r_{(1)}, \ldots, r_{(M)}$, with their corresponding concomitant weights $|v_{[1]}|, \ldots, |v_{[M]}|$.
(3) Sum the concomitant weights starting from $|v_{[1]}|$ and increasing the order.
(4) The weighted median is the sample $r_{(j)}$ whose weight causes the inequality $\sum_{i=1}^{j}|v_{[i]}| \geq v_\varrho$ to hold first.

iii) $p > 1$: In this regime, the minimization of $J_p(\mu)$ corresponds to solving an unconstrained convex optimization problem. Furthermore, since the objective function, $J_p(\mu)$, is twice differentiable and strictly convex for all $\mu \in \mathbb{R}$, we solve (23) using a quasi-Newton method, which does not require the explicit computation of the first- and second-order derivatives at each iteration. Most importantly, the strictly convex nature of $J_p(\mu)$ guarantees convergence to the unique global minimum.

iv) $0 < p < 1$: For these values of p, the minimization of $J_p(\mu)$ corresponds to a nonconvex problem with several local minima. However, it is feasible to find a global minimizer by following a similar to the WMED approach described before. In particular, we first define the fractions $\{r_i = \frac{u_i}{v_i}\}_{i=1}^{M}$, which are then sorted in ascending order, $r_{(1)}, \ldots, r_{(M)}$. Then, the optimization problem in (23) can be reformulated as follows,

$$\mu^{(t)} = \arg\min_{\mu} J_{p,\mathrm{ord}}(\mu) \triangleq \arg\min_{\mu} \sum_{i=1}^{M} |v_{[i]}|^p |\mu - r_{(i)}|^p \,, \tag{26}$$

where $\{|v_{[i]}|\}_{i=1}^{M}$ denote the corresponding concomitant weights. Doing so, the domain of the objective function $J_{p,\mathrm{ord}}(\mu)$ is the union of $M+1$ adjacent intervals, namely, $(-\infty, r_{(1)}]$, $[r_{(1)}, r_{(2)}]$, ..., $[r_{(M-1)}, r_{(M)}]$, $[r_{(M)}, +\infty)$. The critical observation is that, in each interval, the sign of $(\mu - r_{(i)})$, for $i = 1, \ldots, M$, can be determined explicitly. For instance, if $\mu \in [r_{(m-1)}, r_{(m)}]$ for some $1 \leq m \leq M$, then, the objective function becomes

$$J_{p,\mathrm{ord}}(\mu) = \sum_{i=1}^{m-2} |v_{[i]}|^p (\mu - r_{(i)})^p + \sum_{i=m-1}^{M} |v_{[i]}|^p (r_{(i)} - \mu)^p \,. \tag{27}$$

We notice that $(\mu - r_{(i)})^p$ and $(r_{(i)} - \mu)^p$ are all concave functions of μ since $0 < p < 1$. Subsequently, $J_{p,\mathrm{ord}}(\mu)$ is also concave as a nonnegative combination of concave functions. Specifically, $J_{p,\mathrm{ord}}(\mu)$ is piecewise concave in each interval, thus it attains its local minima among the boundary points $\{r_{(i)}\}_{i=1}^{M}$. Finally, since $J_{p,\mathrm{ord}}(\mu) \to +\infty$ as $\mu \to \pm\infty$, we deduce that the global minimizer of (23) is given by

$$\mu^{(t)} = \min_{i=1,\ldots,M} \left\{ J_{p,\mathrm{ord}}(r_{(i)}) \right\} . \tag{28}$$

We note that the proposed rule for updating the step size, as given by (22), guarantees that the ℓ_p^p objective function in (12) does not increase at each iteration. Indeed, the following proposition holds:

Proposition 1 *Let $\mathbf{x}_{\mathcal{S}}^{(t+1)} = \mathbf{x}_{\mathcal{S}}^{(t)} + \mu^{(t)}\mathbf{g}_{\mathcal{S}}^{(t)}$ be the updated sparse solution, where the step size $\mu^{(t)}$ is updated via (22). Then, if $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} = \mathcal{S}$, the proposed step size update guarantees that*

$$\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t+1)}\|_p^p \leq \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t)}\|_p^p \,. \tag{29}$$

Proof: From the optimality of $\mu^{(t)}$ we deduce directly that

$$\begin{aligned} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t+1)}\|_p^p &= \left\| \mathbf{y} - \mathbf{\Phi}\left(\mathbf{x}^{(t)} + \mu^{(t)^*}\mathbf{g}^{(t)}\right) \right\|_p^p \\ &\leq \left\| \mathbf{y} - \mathbf{\Phi}\left(\mathbf{x}^{(t)} + 0\,\mathbf{g}^{(t)}\right) \right\|_p^p = \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t)}\|_p^p \,, \end{aligned}$$

where $\mu^{(t)^*}$ is the optimal step size at iteration t. ■

If the support of $\mathbf{x}^{(t+1)}$ differs from the support of $\mathbf{x}^{(t)}$ estimated at the previous iteration, then the optimality of $\mu^{(t)}$ may not be guaranteed. To alleviate this issue, a backtracking line search is typically used, that is, if

$$\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t+1)}\|_p^p > \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}^{(t)}\|_p^p \,,$$

then, $\mu^{(t)}$ is reduced geometrically, $\mu^{(t)} \leftarrow c_\mu \cdot \mu^{(t)}$, where $c_\mu \in (0,1)$, until the objective function in (12) is reduced. Except if mentioned otherwise, in the subsequent evaluations we set the value of the common ratio equal to $c_\mu = 0.5$.

Furthermore, in order to improve the convergence performance of MD-IHT, a weighting scheme is employed in our implementation, which assigns small weights to large deviations and large weights to small deviations, as they are computed in the previous iteration. Specifically, the suboptimal step size $\mu^{(t)}$ is calculated by setting $\mathbf{u} = \left(\mathbf{W}^{(t)}\right)^{1/2}\left(\mathbf{y} - \mathbf{\Phi}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}^{(t)}\right)$ and $\mathbf{v} = \left(\mathbf{W}^{(t)}\right)^{1/2}\mathbf{\Phi}_{\mathcal{S}}\mathbf{g}_{\mathcal{S}}^{(t)}$ in (23). Doing so, we further suppress the effect of those elements of $\mathbf{x}^{(t)}$ that yield erroneous contributions to the measurements $\mathbf{y}$, while enforcing the contribution of those elements which better agree with the measurements. We also emphasize that, in contrast to the previous reconstruction techniques, our proposed MD-IHT algorithm does not assume any prior information for the sampling noise, such as the scale of its associated distribution.

The algorithm terminates when either a maximum number of iterations, $\mathrm{maxIter}_{\mathrm{MDIHT}}$, has been reached, or the relative change of the ℓ_p^p objective function is less than a predetermined threshold $\mathrm{tol}_{\mathrm{MDIHT}}$. In our implementation we set $\mathrm{maxIter}_{\mathrm{MDIHT}} = 200$ and $\mathrm{tol}_{\mathrm{MDIHT}} = 10^{-16}$. Algorithm 1 summarizes the steps of our MD-IHT sparse reconstruction method.

In the following, we measure the reconstruction quality of the proposed MD-IHT algorithm for the noisy measurements model in (1) by comparing the original sparse signal $\mathbf{x}_0$ with the reconstructed sparse solution $\mathbf{x}_0^*$. Specifically, Theorem 1 below shows that the solution of (12) is an s-sparse signal with an ℓ_2 error that depends on the noise dispersion. This dependence on the noise's pth FLOM, instead of its second-order moment that is either infinite or may not exist, yields an increased robustness of MD-IHT to heavy-tailed sampling noise.

Theorem 1 *Let $\mathbf{x}_0 \in \mathbb{R}^N$ with $\mathcal{S} = \mathrm{supp}(\mathbf{x}_0)$ and $|\mathcal{S}| \leq s$. Assume $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a measurement matrix that satisfies a*

Algorithm 1 The proposed MD-IHT sparse reconstruction algorithm

Input: $\mathbf{y}$, $\mathbf{\Phi}$, s, κ, c_μ, $\mathrm{maxIter}_{\mathrm{MDIHT}}$, $\mathrm{tol}_{\mathrm{MDIHT}}$
Initialize:
SαS parameters: $[\alpha_y, \gamma_y] = \mathrm{Mellinfit}(\mathbf{y})$ $^{(*)}$, $p = \frac{\alpha_y}{2} - 0.001$
Solution: $\mathbf{x}^{(0)} = \mathbf{0}$
Residual: $\mathbf{r}^{(0)} = \mathbf{y}$
Weights: $\mathbf{W}^{(0)}_{i,i} = \frac{1}{\left(\left(r_i^{(0)}\right)^2 + \kappa\right)^{1-\frac{p}{2}}}$, $i = 1, \ldots, M$ (from (17))
Gradient: $\mathbf{g}^{(0)} = -p\, \mathbf{\Phi}^\mathsf{T} \mathbf{W}^{(0)} \mathbf{r}^{(0)}$ (from (14), (15))
$\mathrm{objVal}_{\mathrm{old}} = \|\mathbf{r}^{(0)}\|_p^p$
$\mathcal{S}^{(0)} = \varnothing$, $\mathrm{relChange} = 10^{16}$, $t = 0$

```
while (relChange > tol_MDIHT or t < maxIter_MDIHT) do
    Calculate step size μ^(t) by solving (23) with
    u = (W^(t))^{1/2} (y − Φ_{S^(t)} x^(t)_{S^(t)}) and
    v = (W^(t))^{1/2} Φ_{S^(t)} g^(t)_{S^(t)}
    Update solution via (13): x^(t+1) = H_s (x^(t) + μ^(t) g^(t))
    Update support S^(t+1) = {j | x_j^(t+1) ≠ 0}
    Update residual r^(t+1) = y − Φ_{S^(t+1)} x^(t+1)_{S^(t+1)}
    objVal_new = ||r^(t+1)||_p^p
    {Perform backtracking if necessary}
    if S^(t+1) ≠ S^(t) then
        while objVal_new > objVal_old do
            μ^(t) ← c_μ · μ^(t)
            Update solution via (13):
            x^(t+1) = H_s (x^(t) + μ^(t) g^(t))
            Update support S^(t+1) = {j | x_j^(t+1) ≠ 0}
            Update residual r^(t+1) = y − Φ_{S^(t+1)} x^(t+1)_{S^(t+1)}
            objVal_new = ||r^(t+1)||_p^p
        end while
    end if
    relChange = |objVal_new − objVal_old| / objVal_new
    objVal_old = objVal_new
    Update weights W^(t+1) (from (17))
    Update gradient g^(t+1) = −p Φ^T W^(t+1) r^(t+1) (from (14), (15))
    t = t + 1
end while
```

Output: The original sparse signal $\mathbf{x}_0^* = \mathbf{x}^{(t)}$

$(*)$ $\mathrm{Mellinfit}(\mathbf{y})$ denotes the algorithm described in Section III-D for the estimation of the SαS model parameters from the noisy measurements $\mathbf{y}$ using the method of log-cumulants.

Note: In the general case, the measurement operator $\mathbf{\Phi}$ is replaced by $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}^\mathsf{T}$. Then, MD-IHT estimates a sparse coefficients vector $\boldsymbol{\alpha}_0^*$ and the original signal is obtained by $\mathbf{x}_0^* = \mathbf{\Psi}^\mathsf{T}\boldsymbol{\alpha}_0^*$.

RIP of order $2s$ with $2s$-RIC such that $\delta_{2s} < \sqrt{2} - 1$. Then, for S$\alpha$S sampling noise $\mathbf{n} \in \mathbb{R}^M$ with characteristic exponent α_n and dispersion $\gamma_n \le \epsilon$, the solution of (12), $\mathbf{x}_0^*$, yields a bounded reconstruction error, as follows

$$\|\mathbf{x}_0 - \mathbf{x}_0^*\|_2 \le C_0 \cdot C_{p,\alpha_n} \cdot \epsilon \,, \tag{30}$$

where the constant C_0 depends on δ_{2s}, M and p, and C_{p,α_n} is given by (5).

Proof: Following a similar approach as in [26], we set $\mathbf{x}_0^* = \mathbf{x}_0 + \mathbf{h}$, where $\mathbf{h}$ is a perturbation of the original sparse signal $\mathbf{x}_0$. Since $\mathbf{x}_0^*$ is a feasible point and the dispersion of the error (i.e., the noise) is assumed to be bounded, $\gamma_n \le \epsilon$, it follows that

$$\begin{aligned}\|\mathbf{\Phi}\mathbf{h}\|_2 &= \|\mathbf{\Phi}\mathbf{x}_0^* - \mathbf{\Phi}\mathbf{x}_0\|_2 = \|(\underbrace{\mathbf{\Phi}\mathbf{x}_0^* - \mathbf{y}}_{\mathbf{u}}) + (\underbrace{\mathbf{y} - \mathbf{\Phi}\mathbf{x}_0}_{\mathbf{v}})\|_2 \\ &\overset{(a)}{\le} \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2 \overset{(b)}{\le} \|\mathbf{u}\|_p + \|\mathbf{v}\|_p \le 2\|\mathbf{n}\|_p \\ &\overset{(c)}{\le} 2 \cdot M^{1/p} \cdot (C_{p,\alpha_n} \cdot \gamma_n) \le 2 \cdot M^{1/p} \cdot (C_{p,\alpha_n} \cdot \epsilon) \,,\end{aligned} \tag{31}$$

where (a) follows from the triangle inequality, (b) from the property that if $0 < p < q$ then $\|\mathbf{x}\|_q \le \|\mathbf{x}\|_p$ (in our framework $0 < p < 2$), and (c) is obtained by combining (4) and (7) for the SαS noise $\mathbf{n}$.

To complete the proof, we utilize an intermediate result shown in [26] (see proof of Theorem 1.2), which states that if $\delta_{2s} < \sqrt{2} - 1$ then

$$\|\mathbf{h}\|_2 \le \frac{2\sqrt{1 + \delta_{2s}}}{(1 - \delta_{2s} - \sqrt{2}\delta_{2s})} \|\mathbf{\Phi}\mathbf{h}\|_2 \,. \tag{32}$$

By combining (31) with (32) we have

$$\|\mathbf{h}\|_2 \le \frac{4M^{1/p}\sqrt{1 + \delta_{2s}}}{(1 - \delta_{2s} - \sqrt{2}\delta_{2s})} \cdot (C_{p,\alpha_n} \cdot \epsilon) \,, \tag{33}$$

which is the desired result for $C_0 = \frac{4M^{1/p}\sqrt{1+\delta_{2s}}}{(1-\delta_{2s}-\sqrt{2}\delta_{2s})}$. ■

From (30) we deduce that as the noise dispersion $\gamma_n \to 0$ the reconstruction error approaches zero, whereas in the noiseless case ($\epsilon = 0$) the reconstruction is perfect.

IV. Performance Evaluation

This section evaluates the efficiency of our proposed MD-IHT algorithm as a robust sparse reconstruction method under impulsive sampling noise. To this end, numerical experiments are performed with synthetic signals, along with a comparison against state-of-the-art sparse reconstruction methods tailored to impulsive noise. In particular, the following methods are used for comparisons, which recover the original sparse signal by solving either ℓ_2 or ℓ_p ($p \le 2$) optimization problems: 1) orthogonal matching pursuit (OMP) [1][4], 2) ℓ_p-reweighted least squares (LpRLS) [27][5], and 3) Lorentzian iterative hard thresholding (LIHT) [28]. Appropriate parameters tuning is done for each algorithm according to the guidelines of the associated toolboxes, in order to achieve the optimal reconstruction performance.

The synthetic signals are generated using the following settings, unless stated otherwise: signal length $N = 1024$; cardinality of the sparse support $s = |\mathcal{S}| = \lceil 2\% \cdot N \rceil$; the nonzero coefficients are drawn from a Student's-t distribution with one degree of freedom and their positions are chosen

[4]MATLAB code available from https://goo.gl/VHvyJe.

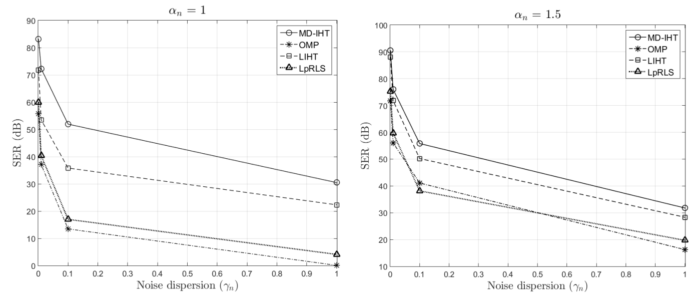
[5]MATLAB code available from https://sites.google.com/site/igorcarron2/cscodes.

uniformly at random from the index set $\{1, 2, \ldots, N\}$; the DCT matrix is used as the sparsifying dictionary $\mathbf{\Psi}$, that is, the measurement operator is given by $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}^{\mathsf{T}}$; the measurement matrix $\mathbf{\Phi}$ has i.i.d. entries drawn from a Bernoulli distribution $\{-1, +1\}$ with equal probability; the number of random projections (measurements) is set to $M = \lceil 25\% \cdot N \rceil$ unless otherwise specified. Furthermore, the results of each experiment are averaged over 500 Monte Carlo repetitions with different realizations of the sparse signals, the random measurement matrices, and the additive noise term. The reconstruction quality is measured in terms of the signal-to-error ratio (SER) (in dB) defined by (9).

Next, the performance of our proposed MD-IHT algorithm is evaluated and compared against OMP, LpRLS, and LIHT. We emphasize again that appropriate tuning is done for the parameters of each algorithm according to the guidelines of the associated toolboxes, in order to achieve the optimal reconstruction performance. For each synthetic signal, a set of linear projections is constructed, which is then corrupted by additive sampling noise following a SαS distribution. For the OMP we assume that the noise tolerance ϵ is known and used as a stopping criterion.

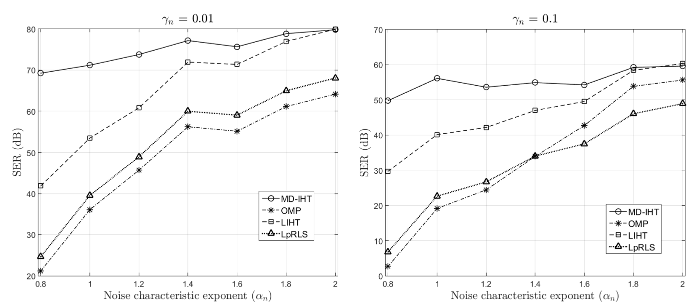
First, we examine the performance of MD-IHT as a function of the noise strength. To this end, we vary the noise dispersion $\gamma_n \in \{0.001, 0.01, 0.1, 1\}$, whilst fixing the noise characteristic exponent $\alpha_n \in \{1, 1.5\}$. The choice of $\alpha_n = 1$ (i.e., Cauchy distribution) is made for a fair comparison with LIHT which is derived from Cauchy statistics. The number of linear projections is set to $M = \lceil 25\% \cdot N \rceil$ and the noise tolerance $\epsilon = M\gamma_n^2$ (for the OMP). Furthermore, all the required SαS parameters in Algorithm 1 are estimated from the noisy measurements directly. Fig. 2 shows the reconstruction performance of each method, in terms of the achieved SER averaged over 500 Monte Carlo runs. As the noise strength increases, the reconstruction accuracy of all methods decreases, as expected. However, MD-IHT yields the highest accuracy among the four methods for the whole range of γ_n and for both the α_n values. Especially in the Cauchy case ($\alpha_n = 1$), this reveals that our MD-IHT algorithm outperforms LIHT, which is tailored to Cauchy statistics, thus demonstrating an increased robustness of MD-IHT to a broader range of noise strength and impulsiveness. Indeed, in contrast to the LIHT whose performance is controlled by tuning only the scale parameter of the Lorentzian norm, the performance of MD-IHT depends on both the value of p in the ℓ_p^p optimization and the estimated dispersion from the noisy measurements. Furthermore, OMP results in the worst reconstruction quality, illustrating the inefficiency of ℓ_2-based methods to suppress the presence of infinite variance sampling noise in the random measurements.

As a second experiment, we evaluate the performance of MD-IHT as a function of the noise impulsiveness. To this end, we vary the noise characteristic exponent $\alpha_n \in [0.8, 2]$, whilst fixing the noise dispersion $\gamma_n \in \{0.01, 0.1\}$. As before, the number of linear projections is set to $M = \lceil 25\% \cdot N \rceil$ and the noise tolerance $\epsilon = M\gamma_n^2$ (for the OMP). Fig. 3 shows the reconstruction performance of each method, in terms of the achieved SER averaged over 500 Monte Carlo runs. As the noise impulsiveness decreases (i.e., $\alpha_n \to 2$) the reconstruction accuracy of all methods increases, whilst they all result in a comparable average SER. As expected, the lower the noise dispersion (γ_n), the higher the reconstruction quality for the four methods. Furthermore, MD-IHT outperforms the other three methods over the whole range of α_n and for both noise dispersion values. Interestingly, both the MD-IHT and the LIHT algorithms, which yield the same performance for Gaussian sampling noise, outperform clearly the least squares-based OMP method, which better adapts to light-tailed environments.



(a) (b)

Fig. 2. Comparison of reconstruction error as a function of noise strength for MD-IHT, OMP, LpRLS, and LIHT. Linear projections are used corrupted by SαS sampling noise with $\alpha_n \in \{1, 1.5\}$ and $\gamma_n \in \{0.001, 0.01, 0.1, 1\}$. Average SER is shown over 500 Monte Carlo runs.



(a) (b)

Fig. 3. Comparison of reconstruction error as a function of noise impulsiveness for MD-IHT, OMP, LpRLS, and LIHT. Linear projections are used corrupted by SαS sampling noise with $\alpha_n \in [0.8, 2]$ and $\gamma_n \in \{0.01, 0.1\}$. Average SER is shown over 500 Monte Carlo runs.

Finally, we examine the reconstruction performance of MD-IHT as the number of linear measurements, M, varies from $2s$ (i.e., twice the cardinality of the sparse support) to $N/2$, for a varying sampling noise impulsiveness with $\alpha_n \in \{0.8, 1.2, 1.6, 2\}$, and a fixed $\gamma_n = 0.05$. Fig. 4 shows that MD-IHT starts yielding fair reconstructions of the original signals using only $M \approx 10\% \cdot N$ corrupted linear measurements. Most importantly, this observation holds even for heavily corrupted measurements (i.e., small α_n values), which illustrates the robustness of MD-IHT in a broad range of impulsive environments. However, as the noise impulsiveness increases more measurements are required to achieve a satisfactory reconstruction quality. This is an expected result, which also resembles the conventional ℓ_2-based reconstruction

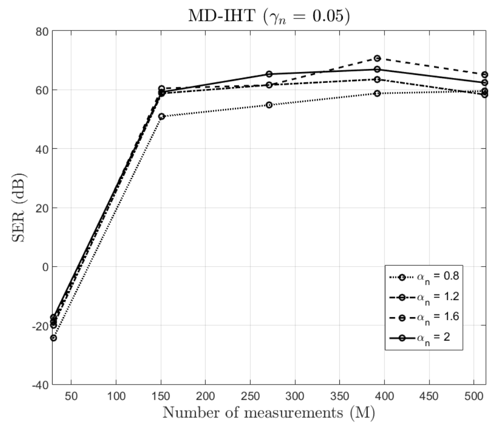


Fig. 4. Average SER of MD-IHT as a function of the number of linear measurements corrupted by SαS sampling noise with $\alpha_n \in \{0.8, 1.2, 1.6, 2\}$ and $\gamma_n = 0.05$.

methods that require more measurements as the noise variance increases.

V. Conclusions and Future Work

In this paper, a robust method was proposed for the reconstruction of sparse signals whose compressive measurements are corrupted by impulsive sampling noise. More specifically, the heavy-tailed statistics of sampling noise with possibly infinite variance was modeled by means of SαS distributions. Subsequently, the effects of additive impulsive sampling noise were suppressed by designing a novel iterative hard thresholding method based on a minimum dispersion (MD) optimization criterion. This criterion emerges naturally in the case of additive sampling noise modeled by SαS distributions. The proposed MD-IHT algorithm demonstrated an increased robustness against gross outliers through a least ℓ_p estimation error criterion, where p depends on the inherent impulsiveness of the noise. A reconstruction error bound was derived that depends on the noise strength, along with rules for tuning the key parameters, such as the value of p and the gradient descent step size, in order to guarantee convergence for a broad range of impulsive noise behaviors. Experimental evaluations revealed that MD-IHT outperforms significantly state-of-the-art methods in the case of highly impulsive sampling noise, whilst resulting in a comparable performance in light-tailed environments.

However, a theoretical framework for selecting the optimal values of the key parameters for the MD-IHT algorithm is still an open question. Furthermore, incorporating some prior knowledge about the unknown sparse support in the reconstruction process typically improves the reconstruction quality. To address this issue, we will examine a modification of the MD-IHT algorithm for stable recovery from compressive measurements given a partially known support.

Acknowledgement

We would like to thank Dr. Rafael Carrillo (CSEM, CH) for providing the MATLAB code and further insights for the LIHT algorithm.

Appendix A
Proof of Positivity of $C_{p,\alpha}^{-p}$

Given the definition in (5), in order to prove the positivity of $C_{p,\alpha}^{-p}$ in (11), it suffices to show that

$$\frac{\Gamma\left(1-\frac{p}{\alpha}\right)}{\cos\left(\frac{\pi}{2}p\right)\Gamma(1-p)} > 0 \ . \qquad (34)$$

Indeed, the requirement for the existence of FLOMs induces that $0 < p < \alpha \leq 2$. This, combined with our empirical rule for setting the optimal value of p as $p \lesssim \frac{\alpha}{2}$ (ref. Section II), yields $0 < p < 1$. Based on this, the sign of each term in (34) is as follows:

i) $p \lesssim \frac{\alpha}{2} \Rightarrow 1-\frac{p}{\alpha} \gtrsim \frac{1}{2} > 0 \Rightarrow \Gamma\left(1-\frac{p}{\alpha}\right) > 0$
ii) $0 < p < 1 \Rightarrow 0 < \frac{\pi}{2}p < \frac{\pi}{2} \Rightarrow 1 > \cos\left(\frac{\pi}{2}p\right) > 0$
iii) $0 < p < 1 \Rightarrow 1 > 1-p > 0 \Rightarrow \Gamma(1-p) > 0$

From i)-iii) we deduce that all the terms are positive, which proves that $C_{p,\alpha}^{-p} > 0$ for $0 < \alpha \leq 2$.

References

[1] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[2] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.
[3] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[4] J. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[5] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.
[6] S. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, Jan. 2010.
[7] X. Tan and J. Li, "Computationally efficient sparse Bayesian learning via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2010–2021, Apr. 2010.
[8] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.
[9] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. and Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[10] F. Liu *et al.*, "Nonconvex compressed sensing by nature-inspired optimization algorithms," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1042–1053, May 2015.
[11] A. Miller, *Subset Selection in Regression*. 2nd Ed., Chapman and Hall/CRC, 2002.
[12] A. Ramirez *et al.*, "Reconstruction of sparse signals from ℓ_1 dimensionality-reduced Cauchy random projections," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5725–5737, Nov. 2012.
[13] R. Carrillo, K. Barner, and T. Aysal, "Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 392–408, Apr. 2010.
[14] G. Samorodnitsky and M. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York, 1994.
[15] G. Tzagkarakis, J. P. Nolan, and P. Tsakalides, "Compressive sensing using symmetric alpha-stable distributions – Part I: Robust nonlinear sampling," *submitted in IEEE Trans. Signal Process.*, Nov. 2017.
[16] E. Candès and P. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, July 2008.

[17] B. Popilka, S. Setzer, and G. Steidl, "Signal recovery from incomplete measurements in the presence of outliers," *Inverse Problems Imag.*, vol. 1, no. 4, pp. 661–672, Nov. 2007.

[18] J. Laska, M. Davenport, and R. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Proc. 43rd Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, 1–4 Nov. 2009.

[19] A. Ramirez, G. Arce, and B. Sadler, "Fast algorithms for reconstruction of sparse signals from Cauchy random projections," in *Proc. 18th Eur. Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, 23–27 Aug. 2010.

[20] G. Tzagkarakis, B. Beferull-Lozano, and P. Tsakalides, "Rotation-invariant texture retrieval with Gaussianized steerable pyramids," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2702–2718, Sept. 2006.

[21] C. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. Wiley, 1995.

[22] M. Davenport *et al.*, "The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4628–4642, Sept. 2012.

[23] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol 27, no. 3, pp. 265–274, Nov. 2009.

[24] A. Achim *et al.*, "Compressive sensing for ultrasound RF echoes using alpha-stable distributions," in *Proc. 32nd Annual Int. Conf. Medicine and Biology Soc. (EMBC)*, Buenos Aires, Argentina, 1–4 Sept., 2010.

[25] Y. Li and G. Arce, "A maximum likelihood approach to least absolute deviation regression," *EURASIP J. Appl. Signal Process.*, vol. 12, pp. 1762–1769, 2004.

[26] E. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, pp. 589–592, 2008.

[27] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, 31 Mar.–4 Apr., 2008.

[28] R. Carrillo and K. Barner, "Lorentzian iterative hard thresholding: Robust compressed sensing with prior information," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4822–4833, Oct. 2013.