Dictionary Learning with Statistical Sparsity in the Presence of Noise

Shayan Aziznejad
Biomedical Imaging Group
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
shayan.aziznejad@epfl.ch

Emmanuel Soubies
IRIT, CNRS
Université de Toulouse
France
emmanuel.soubies@irit.fr

Michael Unser
Biomedical Imaging Group
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
michael.unser@epfl.ch

***Abstract*—We consider a new stochastic formulation of sparse representations that is based on the family of symmetric α-stable (SαS) distributions. Within this framework, we develop a novel dictionary-learning algorithm that involves a new estimation technique based on the empirical characteristic function. It finds the unknown parameters of an SαS law from a set of its noisy samples. We assess the robustness of our algorithm with numerical examples.**

***Index Terms*—Dictionary learning, sparse coding, sparse representation, stable distribution, empirical characteristic function.**

I. Introduction

The sparse representation of data is a fundamental approach to modern signal processing; it has been extensively developed in the past 20 years [1]. This approach appears in numerous research areas such as compressed sensing [2], inverse problems [3], and image processing [4], to name a few. In this framework, the data vector $\mathbf{y} \in \mathbb{R}^M$ is assumed to have a sparse representation in the span of the columns (atoms) of some transformation matrix $\mathbf{A} \in \mathbb{R}^{M\times P}$. In other words, one can write that

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1}$$

for some sparse vector $\mathbf{x} \in \mathbb{R}^P$ with few nonzero entries. Finding the sparsest representation of $\mathbf{y}$ is then formulated as the minimization

$$\min_{\mathbf{x}\in\mathbb{R}^P} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \tag{2}$$

where $\|\mathbf{x}\|_0$ is the number of nonzero elements of $\mathbf{x}$. Problem (2) is known to be NP-hard. However, there is a rich literature on how to efficiently find (or approximate) the solution of (2) (see [5] for a review).

A prominent element in this framework is the transformation matrix (a.k.a. dictionary) $\mathbf{A}$. Choosing a suitable dictionary is a challenging task and highly depends on the application. For example, in compressed sensing, Gaussian random matrices are advantageous since they ensure stable recovery by satisfying the restricted isometry property [6]. Meanwhile, in image-compression algorithms, dictionaries such as the discrete cosine tranform [7] or the various wavelet transforms [8] are classical choices since they provide a sparse (hence, compressible) representation for natural images. As an alternative to the classical approach, the framework of dictionary learning has been proposed. In this data-driven scheme, the transformation itself is learned from a set of data samples [9].

While classical sparse models are purely deterministic, the theory of sparse stochastic processes has been developed in the last decade to model sparsity in the stochastic world [10]. Within this framework, the class of sparsity-promoting distributions has been well characterized [11]. More precisely, it has been shown that the realizations of a vector of i.i.d. random variables with heavy-tailed distributions are sparse [12]. Symmetric α-stable (SαS) distributions are a popular family of heavy-tailed distributions which are appearing in many research areas, for example in compressed sensing [13], audio denoising [14] and modeling of financial data [15]. We refer to [16] for a list of applications in signal processing.

In this paper, we consider dictionary learning with a statistical prior according to which the components of the signal $\mathbf{x}$ are i.i.d. with a SαS law. In the noiseless scenario, Pad *et al.* proposed a novel algorithm [17] that we adapt here to the more realistic case where an additive Gaussian noise is assumed to corrupt the data samples. Our method is based on a robust estimation of the noise variance jointly with the parameters of the underlying SαS distribution. We numerically illustrate the performance of our algorithm and verify its robustness to noise.

II. Preliminaries and Model

In this section, we first recall some important properties of SαS distributions. Then, we detail the stochastic formulation of sparse representation that we are considering throughout the paper.

A. Symmetric α-Stable Distributions

The SαS random variable X is defined via its characteristic function given as

$$\Phi_X(\omega) = \mathbb{E}[\mathrm{e}^{\mathrm{j}\omega X}] = \exp(-\gamma|\omega|^\alpha), \tag{3}$$

where $\gamma > 0$ and $\alpha \in (0, 2]$ are the dispersion and stability parameters of X, respectively. The extremal case $\alpha = 2$ coincides with the zero-mean Gaussian law whose variance is $\sigma^2 = 2\gamma$.

SαS distributions are closed under addition, in the sense that any linear combination of independent α-stable random

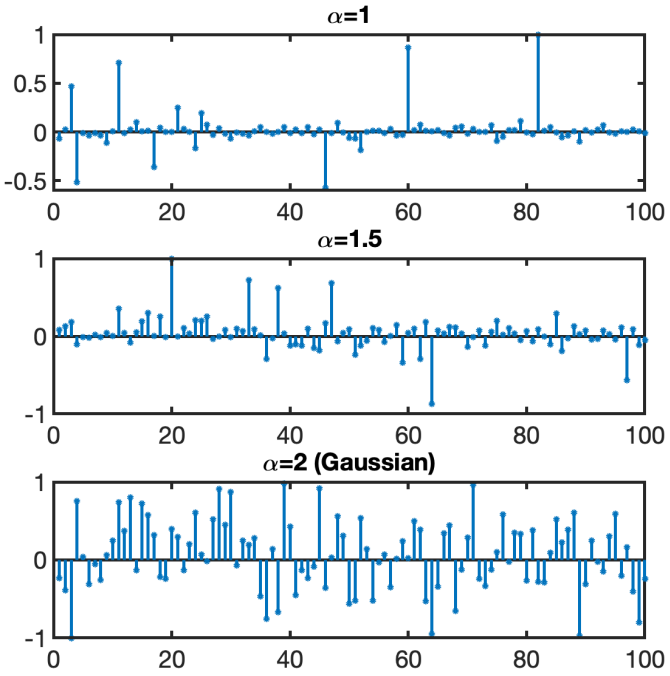

Fig. 1. Three realizations of SαS signals with different values of α. Each signal has been divided by its max value.

variables is an α-stable random variable itself [18]. More precisely, if $\{X_n\}_n^N$ is an i.i.d. sequence of random variables with SαS law, then the characteristic function of $X = \sum_{n=1}^{N} w_n X_n$ can be computed as

$$\Phi_X(\omega) = \exp(-\gamma \|\mathbf{w}\|_\alpha^\alpha |\omega|^\alpha), \tag{4}$$

where $\mathbf{w} = (w_1, \ldots, w_N) \in \mathbb{R}^N$.

For small values of α, it is known that the independent realizations of a random variable with SαS law yields a sparse signal [12]. As α increases, the sparsity of the signal decreases. In the critical case $\alpha = 2$ (zero-mean Gaussian distribution), the realizations are purely non sparse. We illustrate this behavior in Figure 1 where, in each subplot, 100 independent samples of an SαS law with $\gamma = 1$ and with different values of α have been generated.

B. Model

We consider that the data vector $\mathbf{y}$ follows the model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{5}$$

where $\mathbf{x} = (x_1, \ldots, x_P) \in \mathbb{R}^P$ is a realization of a vector of P i.i.d. random variables that follow an SαS distribution, while $\mathbf{n} \in \mathbb{R}^M$ is an additive white Gaussian noise (independent of $\mathbf{x}$) with the variance σ^2. This model is similar to the classical sparse-representation scheme (1) with two major differences.

- It considers additive Gaussian noise, which makes the model more realistic.
- The prior information on the latent signal $\mathbf{x}$ is stochastic, being a realization of a random vector with SαS law.

We assume that the stable law that generates the vector $\mathbf{x}$ has dispersion parameter $\gamma = 1$. This covers in no loss of generality, since γ can be absorbed as a scaling factor in the columns of $\mathbf{A}$.

Now, the aim of dictionary learning in this framework is to learn the unknown matrix $\mathbf{A}$ from the training dataset $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$. This dataset is obtained by taking independent instances from the stochastic model (5).

III. Sparse-Distribution Tomography

In this section, we briefly explain the sparse distribution tomography (SparsDT) algorithm of Pad *et al.* for learning the matrix $\mathbf{A}$ in the model (5) [17].

For any unit-norm vector $\mathbf{u} \in \mathbb{R}^M$ with $\|\mathbf{u}\|_2 = 1$, the random variable

$$Z_\mathbf{u} = \mathbf{u}^T\mathbf{y} = \mathbf{u}^T\mathbf{A}\mathbf{x} \tag{6}$$

follows a SαS distribution. Its characteristic function is

$$\Phi_{Z_\mathbf{u}}(\omega) = \exp\left(-\|\mathbf{A}^T\mathbf{u}\|_\alpha^\alpha |\omega|^\alpha\right). \tag{7}$$

The set $\{z_1, \ldots, z_K\}$ of sample points of $Z_\mathbf{u}$, with $z_k = \mathbf{u}^T\mathbf{y}$, allows one to estimate the parameters α (stability) and $\gamma(\mathbf{u}) = \|\mathbf{A}^T\mathbf{u}\|_\alpha^\alpha$ (dispersion) of the law of $Z_\mathbf{u}$. By repeating this procedure for the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_L\}$, one gets the system of nonlinear equations

$$\|\mathbf{A}^T\mathbf{u}_\ell\|_\alpha^\alpha = \gamma(\mathbf{u}_\ell), \quad \ell = 1, \ldots, L, \tag{8}$$

for the unknown dictionary $\mathbf{A}$. It was shown in [17] that, if L is sufficiently large (namely, $L \geq MP$), then the solution of (8) is unique (up to negation and permutations). Hence, by solving the nonlinear system of equations (8), one recovers the desired dictionary $\mathbf{A}$.

Although SparsDT is an elegant algorithm that works perfectly in noiseless (or very low noise) regimes, we observe that its performance decreases dramatically as the noise power increases. This is due mainly to their estimation method for finding the unknown parameters of the law of $Z_\mathbf{u}$ from its samples. In the next section, we propose an alternative estimation method that takes additive Gaussian noise into account and results in a robust version of SparsDT with improved performance in high-noise regimes.

IV. Parameter Estimation Using Empirical Characteristic Functions

Consider the random variable $Z = X + N$, where X follows a SαS law with the dispersion parameter $\gamma > 0$ and N is a zero-mean Gaussian random variable that does not depend on X and has the variance σ^2. In this section, we propose a new method to estimate the parameters $\alpha \in (0, 2]$, $\gamma > 0$, and $\sigma > 0$ of this model from the set $\{z_1, \ldots, z_K\}$ of independent samples of Z.

We start by forming the empirical characteristic function

$$\tilde{\Phi}_Z(\omega) = \frac{1}{K}\sum_{k=1}^{K} \exp(\mathrm{j}\omega z_k). \tag{9}$$

From the generic form (3) of characteristic functions of SαS random variables and together with the independence of X and Z, one readily verifies that the characteristic function of Z is indeed

$$\Phi_Z(\omega) = \exp\left(-\gamma|\omega|^\alpha - \frac{\sigma^2}{2}\omega^2\right). \tag{10}$$

Now, the law of large numbers suggests that, for large values of K, the empirical characteristic function of Z should be close to its theoretical form (10). So, in order to estimate the the unknown parameters of Z, we fit (10) to (9). To that end, we define the vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_T)$ of modulation samples and the function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ with $\tilde{f}(\omega) = -\log(\tilde{\Phi}_Z(\omega))$ for all $\omega \in \mathbb{R}$. After fitting (10) to (9), we want to have that

$$\tilde{f}(w_t) \approx \gamma|\omega_t|^\alpha + \frac{\sigma^2}{2}\omega_t^2, \quad t = 1, \ldots, T.$$

Consequently, in order to find the unknown parameters of Z, we propose to solve the least-squares minimization

$$\min_{\substack{\alpha\in(0,2]\\ \gamma,\sigma\geq 0}} \sum_{t=1}^{T}\left(\tilde{f}(\omega_t) - \gamma|\omega_t|^\alpha - \frac{\sigma^2}{2}\omega_t^2\right)^2. \tag{11}$$

We first note that, by defining the intermediate cost function

$$\mathcal{J}(\alpha) = \min_{c_1,c_2>0} \sum_{t=1}^{T}\left(\tilde{f}(w_t) - c_1|\omega_t|^\alpha - c_2\omega_t^2\right)^2, \tag{12}$$

Problem (11) is equivalent to

$$\min_{\alpha\in(0,2]} \mathcal{J}(\alpha), \tag{13}$$

with the implicit change of variables $c_1 = \gamma$ and $c_2 = \sigma^2/2$. The promising property of this decoupling is that (12) is a simple linear least-squares problem that is known to have a unique solution with closed form. By writing the optimality conditions (partial derivatives of the cost function equal to 0) for (12), we obtain that the solution pair $\mathbf{c}^*(\alpha) = (c_1^*(\alpha), c_2^*(\alpha))$ of (12) must satisfy

$$\sum_{t=1}^{T} |w_t|^\alpha(\tilde{f}(w_t) - c_1|w_t|^\alpha - c_2 w_t^2) = 0$$
$$\sum_{t=1}^{T} |w_t|^2(\tilde{f}(w_t) - c_1|w_t|^\alpha - c_2 w_t^2) = 0.$$

Solving this system of linear equations then yields

$$\mathbf{c}^*(\alpha) = \begin{pmatrix} \|\boldsymbol{\omega}\|_{2\alpha}^{2\alpha} & \|\boldsymbol{\omega}\|_{2+\alpha}^{2+\alpha} \\ \|\boldsymbol{\omega}\|_{2+\alpha}^{2+\alpha} & \|\boldsymbol{\omega}\|_4^4 \end{pmatrix}^{-1} \begin{pmatrix} g_\alpha \\ g_2 \end{pmatrix}, \tag{14}$$

where $g_p = \sum_{t=1}^{T} |\omega_t|^p \tilde{f}(\omega_t)$ for $p = \alpha, 2$. The conclusion is that, for each value of α, there is a fast way of computing the cost function $\mathcal{J}(\alpha)$ and also the (respectively) optimal coefficients $\gamma = c_1^*(\alpha)$ and $\sigma = \sqrt{2c_2^*(\alpha)}$.

The final step is to to find the stability parameter α that minmizes (13). We recall that α lies in the finite-length interval $(0, 2]$. Hence, we can benefit from the rich existing liteature on derivative-free methods for the optimization of one-dimensional functions over compact domains (*e.g.*, Bayesian optimization [19]) to solve (13).

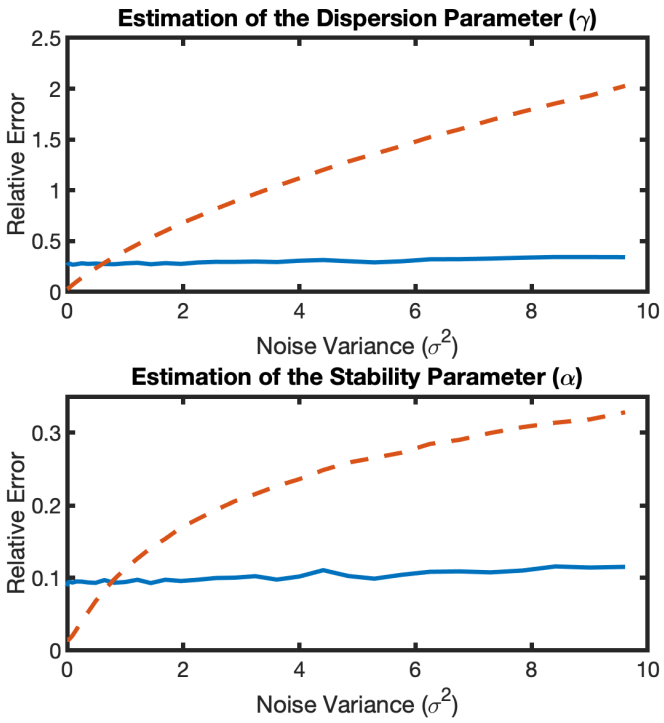

Fig. 2. Estimation of the stability (α) and the dispersion (γ) parameters of an SαS random variable from a noisy sample set of size $K = 10000$. We compare our proposed method (solid line) to the one of SparsDT (dashed line). An average over 1000 repetitions has been taken to provide a smooth curve.

In Figure 2, we show that even a simple grid search with stepsize $h = 0.01$ and $\boldsymbol{\omega} = (0.1, 0.11, 0.12, \ldots, 0.3)$ provides a suitable estimation of stability parameter as $\alpha = 1.255$ and the dispersion parameter $\gamma = 1$.

V. Numerical Results

In this section, we compare the robustness of our method with SparsDT through numerical examples. To that end, we first generated a random Gaussian matrix $\mathbf{A}$ of size $M = 20$ by $P = 30$. Then, we generated a training dataset $\{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ of size $K = 1000$, where each data vector is an independent realization of (5) with $\alpha = 1.2$ and $\gamma = 1$ and with different noise levels (we swipe the variance of the noise over the range $[0, 10]$).

To measure the performance of dictionary-learning methods, we use the average correlation metric that was introduced in [17]. In this metric, the distance between the learned dictionary $\widehat{\mathbf{A}}$ and the ground truth $\mathbf{A}$ is obtained in two stages; first, we pair each column $\widehat{\mathbf{a}}_n$ of $\widehat{\mathbf{A}}$ to an unpaired column of $\mathbf{A}$ (one-to-one assignment) that is maximally correlated to $\widehat{\mathbf{a}}_n$. Then, the distance is computed by taking an average over the correlation of these M pairs.

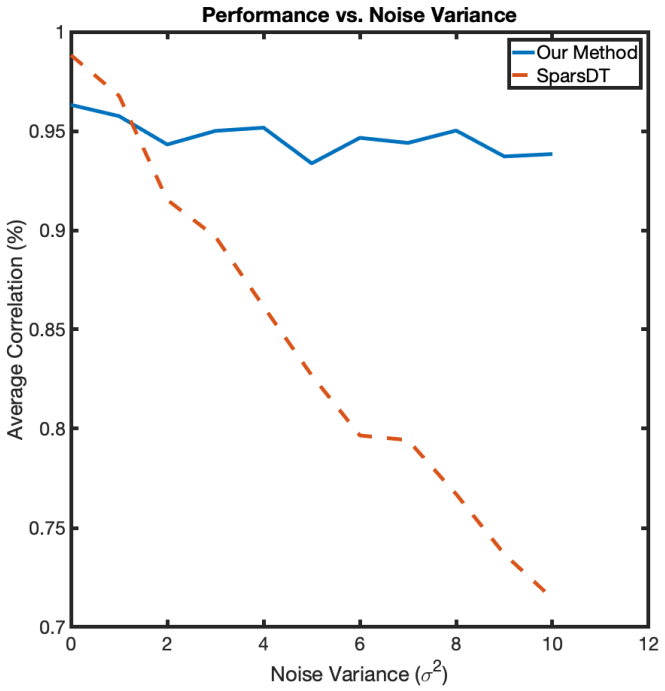

Fig. 3. Effect of noise variance (σ^2) to the performance of the SparsDT algorithm (dashed line) and our modified version (solid line).

We provide our results in Figure 3. As can be seen, our method is robust to noise. Indeed, in low-noise regimes, SparsDT provides a better performance due to its accurate estimate of the parameters of SαS laws. But, the SparsDT approach is not robust to an increase in the noise variance. By contrast, our method has a better performance (correlation around 95 percent) even in high-noise regimes.

VI. Acknowledgement

This work was supported in part by the Swiss National Science Foundation under Grant 200020_184646 / 1 and in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 692726-GlobalBioIm.

VII. Conclusion

We have proposed a dictionary-learning method with a statistical sparsity prior in the presence of noise. We have assumed that the sparse signal is a realization of an i.i.d. vector whose entries obey a symmetric α-stable (SαS) law. We have proposed a modification to an existing algorithm that made it robust to the noise variance. Our modification is based on the estimation of the dispersion parameter of a SαS law from a set of its noisy samples. We have validated numerically the robustness of our approach. In our future works, we plan to investigate different heavy-tailed distributions as statistical sparsity prior and also to apply our method in the real-world problems that follow the SαS model.

References

[1] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer Science & Business Media, 2010.

[2] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[4] M. Elad, M. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[5] F. Marvasti, A. Amini, F. Haddadi, M. Soltanolkotabi, B. Khalaj, A. Aldroubi, S. Sanei, and J. Chambers, "A unified approach to sparse signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 44, 2012.

[6] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[7] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[8] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

[9] I. Tosic and P. Frossard, "Dictionary learning: What is the right representation for my signal?," *IEEE Signal Processing Magazine*, vol. 28, pp. 27–38, 2011.

[10] M. Unser and P.D. Tafti, *An Introduction to Sparse Stochastic Processes*, Cambridge University Press, Cambridge, United Kingdom, 2014, 367 p.

[11] A. Amini and M. Unser, "Sparsity and infinite divisibility," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2346–2358, 2014.

[12] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5193–5201, 2011.

[13] George Tzagkarakis, John P Nolan, and Panagiotis Tsakalides, "Compressive sensing using symmetric alpha-stable distributions for robust sparse signal reconstruction," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 808–820, 2018.

[14] N. Bassiou, C. Kotropoulos, and E. Koliopoulou, "Symmetric α-stable sparse linear regression for musical audio denoising," in *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2013, pp. 382–387.

[15] J. Nolan, *Stable Distributions: Models for Heavy-Tailed Data*, Birkhauser Boston, 2003.

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

[17] P. Pad, F. Salehi, E. Celis, P. Thiran, and M. Unser, "Dictionary learning based on sparse distribution tomography," in *Proceedings of the Thirty-Fourth International Conference on Machine Learning (ICML'17)*, Sydney, Commonwealth of Australia, August 6-11, 2017, pp. 2731–2740.

[18] G. Samoradnitsky and M.S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, New York, 2017, 632 p.

[19] J. Mockus and L. Mockus, "Bayesian approach to global optimization and application to multiobjective and constrained problems," *Journal of Optimization Theory and Applications*, vol. 70, no. 1, pp. 157–172, 1991.