Optimality of Operator-Like Wavelets for Representing Sparse AR(1) Processes

Pedram Pad, *Student Member, IEEE*, and Michael Unser, *Fellow, IEEE*

***Abstract*—The discrete cosine transform (DCT) is known to be asymptotically equivalent to the Karhunen-Loève transform (KLT) of *Gaussian* first-order auto-regressive (AR(1)) processes. Since being uncorrelated under the Gaussian hypothesis is synonymous with independence, it also yields an independent-component analysis (ICA) of such signals. In this paper, we present a constructive non-Gaussian generalization of this result: the characterization of the optimal orthogonal transform (ICA) for the family of symmetric-α-stable AR(1) processes. The degree of sparsity of these processes is controlled by the stability parameter $0 < \alpha \leq 2$ with the only non-sparse member of the family being the classical Gaussian AR(1) process with $\alpha = 2$. Specifically, we prove that, for $\alpha < 2$, a fixed family of operator-like wavelet bases systematically outperforms the DCT in terms of compression and denoising ability. The effect is quantified with the help of two performance criteria (one based on the Kullback-Leibler divergence, and the other on Stein's formula for the minimum estimation error) that can also be viewed as statistical measures of independence. Finally, we observe that, for the sparser kind of processes with $0 < \alpha \leq 1$, the operator-like wavelet basis, as dictated by linear system theory, is undistinguishable from the ICA solution obtained through numerical optimization. Our framework offers a unified view that encompasses sinusoidal transforms such as the DCT and a family of orthogonal Haar-like wavelets that is linked analytically to the underlying signal model.**

***Index Terms*—Operator-like wavelets, independent-component analysis, auto-regressive processes, stable distributions.**

I. Introduction

TRANSFORM-DOMAIN processing is a classical approach to compress signals, model data, and extract features. The guiding principle is to produce transform-domain coefficients that are decoupled statistically so that a simple component-wise processing can be applied; i.e., each coefficient is processed independently of the others. The reference solution in the field is the Karhunen-Loève transform (KLT) which yields transform-coefficients that are uncorrelated and therefore also independent, provided the process is Gaussian. Also, if the process is stationary with finite variance and infinite length, then the KLT is a Fourier-like transform [1]. Moreover, it has been shown that the discrete cosine transform (DCT) [2] is asymptotically equivalent to the KLT for the whole class of stationary processes [3], including the AR(1) model [4]; thus, for a Gaussian input, all these transforms result in a fully decoupled (independent) representation. However, this favorable independence-related property is extinguished for non-Gaussian processes. In this case, the coefficients are only partially decoupled and the representation of the signal afforded by the KLT is no longer optimal.

In recent years, wavelets have emerged as an alternative representation of signals and images. Typical examples of successful applications are JPEG2000 for image compression [5] and shrinkage methods for attenuating noise [6], [7]. The fact that wavelets are so effective in transform-domain applications suggests that they are naturally suited to represent practical processes. This empirical observation was established by early studies that include [8], where many natural images were subjected to an independent-component analysis (ICA). It was found that the resulting components have properties that are reminiscent of 2D wavelets and/or Gabor functions. Additional ICA experiments were performed in [9] on realizations of the stationary sawtooth process and of Meyer's ramp process [10]; for both processes, the basis vectors of ICA exhibit a wavelet-like multiresolution structure.

Despite their empirical usefulness, the optimality of wavelets for the representation of non-Gaussian stochastic processes remains poorly understood from a theoretical point of view. An early study can be traced back to [11], where the decomposition of fractional Brownian motions over a wavelet basis was shown to result in almost uncorrelated coefficients, under some conditions. By contrast, in the deterministic framework, it is well known that wavelets are optimal (up to some constant) for the N-term approximation of functions in Besov spaces [12]; the extension of this result to a statistical setting could be achieved only experimentally.

Recently, a general distributional framework for the specification of sparse stochastic processes has been proposed in [13], [14]. It is particularly well suited to the specification of symmetric-α-stable (SαS) white noises, which can be used to drive first-order stochastic differential equations (SDE) to synthesize AR(1) processes. As it turns out, AR(1) systems and α-stable distributions are at the core of signal modeling and probability theory. The classical Gaussian processes correspond to $\alpha = 2$, while $0 < \alpha < 2$ yields stable processes that have heavy-tailed statistics and that are prototypical representatives for sparse signals [15], [16]. Such stable models are attractive for statistical signal processing because they lend themselves well to analytic calculations [17]. Areas of applications include detection theory [18], communications [19], and signal denoising [20]. Also, specifically heavy-tailed AR have been used to model

Manuscript received August 19, 2014; revised February 18, 2015 and May 26, 2015; accepted June 06, 2015. Date of publication June 19, 2015; date of current version August 13, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ruixin Niu. This work was supported by the European Commission under Grant ERC-2010-AdG 267439-FUN-SP.

The authors are with the Biomedical Imaging Group, EPFL, Lausanne CH-1015, Switzerland (e-mail: pedram.pad@epfl.ch; michael.unser@epfl.ch).

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

Digital Object Identifier 10.1109/TSP.2015.2447494

phenomena in network [21], sea surface [22], economy and finance [23].

In this paper, we take advantage of this framework to establish the optimality of a certain class of wavelets in a stochastic sense. We start by characterizing the amount of dependency between the coefficients of stochastic processes represented in an arbitrary transform domain. To that end, we introduce two performance criteria. The first assesses the coding performance of the transform: it is given by the Kullback-Leibler divergence between the joint probability density function (pdf) of the original signal and the product of the marginals in the transformed domain. The second is a theoretical prediction of denoising performance under the hypothesis of additive white Gaussian noise (AWGN). It is based on Stein's formula for the mean-square estimation error and also takes the form of a divergence between the joint pdf of the original signal and the product of the marginals in the transformed domain. Then, we seek the orthogonal transformation that minimizes these statistical criteria. We confirm the loss of optimality of the DCT for $0 < \alpha < 2$ and validate the superiority of a special brand of operator-like wavelet transform that is matched to the underlying signal model. Our reference method in this comparison is the ICA solution that is determined by numerical means for different values of α. The remarkable empirical finding of this paper is that the ICA solution converges to the operator-like wavelets for values of α below one.

The practical relevance of these results is that, unlike ICA, the operator-like wavelets are known in analytical form in terms of the pole of the underlying system (see (29)). They are a special case of the differential wavelets investigated in [24]. They may also be interpreted as a generalization of the Haar transform with scale-dependent filters. In essence, this amounts to replacing the finite-difference operations of the conventional wavelet transform algorithm by a suitable series of linear prediction errors where the coefficients are determined by the pole of the AR(1) system.

This paper is organized as follows: In Section II, we introduce two measures of divergence between distributions that are suitable for either noise attenuation or compression applications. The signal model fundamental to this paper is discussed in Section III.A and III.B. The operator-like wavelets that are deduced from this model are presented in Section III.C. In Section IV, we derive the explicit form of our performance criteria for the SαS model in the context of transform-domain compression and noise attenuation. In addition, we provide an iterative algorithm to find the optimal basis. Results for different AR(1) processes and different transform domains are discussed in Section V. The last section is dedicated to the recapitulation of the main results, the relation to prior works, and topics for future studies.

II. Performance Measures

In statistical signal processing, it is of interest to precisely quantify the best-achievable performance when the model is not perfectly matched to the signal under investigation, or when certain simplifying hypothesis, such as independence, are being made. In the following, we address this issue for the two problems of compression and denoising when the assumed distribution and the real one may differ.

1) Compression Based on Non-Exact Distribution: It is well-known that, if we have a source $\mathbf{s}$ of iid random vectors with common pdf $p_{\mathbf{s}}$, then the logarithm of measure of the coding set per sample can be at least

$$\mathbb{H}(p_{\mathbf{s}}) = -\int p_{\mathbf{s}}(\mathbf{s}) \log p_{\mathbf{s}}(\mathbf{s}) \mathrm{d}\mathbf{s} \tag{1}$$

which is the entropy[1] of the source [25]. However, if we compress $\mathbf{s}$ assuming that it is distributed according to $q_{\mathbf{s}}$ (rather than $p_{\mathbf{s}}$), then

$$\begin{aligned} \mathbb{H}(q_{\mathbf{s}}) &= \mathbb{H}(p_{\mathbf{s}}) + \mathbb{D}(p_{\mathbf{s}} \| q_{\mathbf{s}}) \\ &= \mathbb{H}(p_{\mathbf{s}}) + \int p_{\mathbf{s}}(\mathbf{s}) \log \frac{p_{\mathbf{s}}(\mathbf{s})}{q_{\mathbf{s}}(\mathbf{s})} \mathrm{d}\mathbf{s} \end{aligned} \tag{2}$$

in which $\mathbb{D}(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

Typically, when there is a statistical dependency between the entries of $\mathbf{s}$, compressing the vector based on the exact distribution is often intractable. Thus, the common strategy is to expand the vector in some other basis and to then do the compression entry-wise (neglecting the dependency between entries of the transformed vector). This is equivalent to doing the compression assuming that the signal distribution is the product of the marginal distributions. Thus, if the transformed vector is $\mathbf{y} = \mathbf{H}\mathbf{s}$, then the normalized redundant information remaining in the compressed signal is

$$\begin{aligned} \mathrm{R}(\mathbf{H}) &= \frac{1}{N} \left(\mathbb{H}(p_{y_1}(y_1) \cdots p_{y_N}(y_N)) - \mathbb{H}(p_{\mathbf{s}}) \right) \\ &= \frac{1}{N} \mathbb{D}\left(p_{\mathbf{y}}(\mathbf{y}) \| p_{y_1}(y_1) \cdots p_{y_N}(y_N) \right), \end{aligned} \tag{3}$$

where N is the number of entries in $\mathbf{s}$. This is the first measure of performance of the transform $\mathbf{H}$ that we use in this paper. Also, this criterion is commonly used in ICA to find the "most-independent" representation [26].

2) Denoising Based on Non-Exact Distribution: Although the Kullback-Leibler divergence is widely used to measure the distance between two distributions, it is inherently tied to the application of compression. Here, we introduce a novel measure of divergence between distributions that is more specifically targeted to the classical denoising task. Consider the problem of estimating $\mathbf{s}$ from the noisy measurement

$$\mathbf{z} = \mathbf{s} + \mathbf{n} \tag{4}$$

where $\mathbf{n}$ is an N-dimensional Gaussian random vector with iid entries with variance σ^2 that is also independent from $\mathbf{s}$. Our prior knowledge is the Nth order pdf $p_{\mathbf{s}}(\cdot)$ of the signal. Under these assumptions and according to Stein [27], the optimal signal estimator that obtains minimum mean-square error (MMSE) is

$$\mathbb{E}\{\mathbf{s}|\mathbf{z}\} = \mathbf{z} + \sigma^2 \boldsymbol{\nabla} \log p_{\mathbf{z}}(\mathbf{z}) \tag{5}$$

where $\mathbb{E}\{\mathbf{s}|\mathbf{z}\}$ is the expected value of $\mathbf{s}$ given $\mathbf{z}$, $p_{\mathbf{z}}(\mathbf{z}) = (p_{\mathbf{s}} * p_{\mathbf{n}})(\mathbf{z})$ is the Nth order pdf of the noisy measurements, and $\boldsymbol{\nabla}$ represents the gradient operator. Thus, the MSE given $\mathbf{z}$ is

$$\begin{aligned} &\mathbb{E}\left\{ (\mathbf{s} - \mathbb{E}\{\mathbf{s}|\mathbf{z}\})^2 | \mathbf{z} \right\} \\ &= \int \|\mathbf{s} - \mathbf{z}\|^2 p(\mathbf{s}|\mathbf{z}) \mathrm{d}\mathbf{s} - \sigma^4 \|\boldsymbol{\nabla} \log p_{\mathbf{z}}(\mathbf{z})\|^2 \\ &= N\sigma^2 + \sigma^4 \Delta \log p_{\mathbf{z}}(\mathbf{z}). \end{aligned} \tag{6}$$

[1] $\mathbb{H}(\cdot)$ is used for the random variable or its pdf interchangeably.

where Δ is the Laplacian operator. Averaging over $\mathbf{z}$, we have

$$\begin{aligned}\mathrm{MMSE} &= N\sigma^2 - \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \|\boldsymbol{\nabla} \log p_{\mathbf{z}}(\mathbf{z})\|^2 \mathrm{d}\mathbf{z} \\ &= N\sigma^2 + \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \Delta \log p_{\mathbf{z}}(\mathbf{z}) \mathrm{d}\mathbf{z}, \end{aligned} \tag{7}$$

However, if we apply this signal estimator based on an incorrect prior $q_{\mathbf{s}}$ (instead of the true distribution $p_{\mathbf{s}}$) as the distribution of $\mathbf{s}$, then by using (5)–(7), the MSE of estimation becomes

$$\mathrm{MSE}(q_{\mathbf{s}}) = \mathrm{MMSE} + \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \left\| \boldsymbol{\nabla} \log \frac{p_{\mathbf{z}}(\mathbf{z})}{q_{\mathbf{z}}(\mathbf{z})} \right\|^2 \mathrm{d}\mathbf{z} \tag{8}$$

where $q_{\mathbf{z}}(\mathbf{z})$ is the distribution induced on $\mathbf{z}$ in (4) when the distribution on $\mathbf{s}$ is $q_{\mathbf{s}}(\mathbf{s})$. Here, notice the pleasing similarity between (1)–(2) and (7)–(8).

If the entries of $\mathbf{s}$ are dependent, then the entries of $\mathbf{z}$ are dependent, too. Then, performing the exact MMSE estimator is once again often infeasible. The common scheme is then to take $\mathbf{z}$ into a transform domain, perform an entry-wise denoising (regardless of the dependency between coefficients), and map the result back into the original domain. This is justifiable when the transformation $\mathbf{H}$ is unitary because the transform-domain noise remains Gaussian iid while the ℓ_2-norm of the signal is preserved. Hence, the expected performance of this scalar denoising scheme is $\mathrm{MSE}(p_{\tilde{y}_1}(\tilde{y}_1) \cdots p_{\tilde{y}_N}(\tilde{y}_N))$ where $p_{\tilde{y}_n}(\tilde{y}_n)$ is the marginal distribution of the nth entry of $\tilde{\mathbf{y}} = \mathbf{H}\mathbf{z}$. We write this as a function of $\mathbf{H}$ normalized by the dimensionality of $\mathbf{s}$, with

$$\mathrm{MSE}(\mathbf{H}) = \frac{1}{N} \mathrm{MSE}\left(p_{\tilde{y}_1}(\tilde{y}_1) \cdots p_{\tilde{y}_N}(\tilde{y}_N)\right) \tag{9}$$

which is the second measure of performance that we consider in this paper.

III. Modeling and Wavelet Analysis of SαS AR(1) Processes

In this section, we present a continuous-domain description of a SαS AR(1) process as the solution of a first-order stochastic differential equation. This differential formulation is central to our argumentation since it results in the identification of the operator-like wavelets, as discussed in Section III.C. We also show that the continuous-domain representation is consistent with the more standard discrete AR(1) model in the sense that the latter is the sampled version of the former.

A. Differential Modeling of SαS AR(1) Processes

In [13], the authors define a sparse stochastic process s as the solution of the linear differential equation

$$\mathrm{L}s = w \tag{10}$$

where L is a suitable differential operator and w a non-Gaussian continuous-domain white noise (or innovation process). Formally, this results into the solution

$$s = \mathrm{L}^{-1} w \tag{11}$$

where the linear operator L^{-1} is the inverse of the whitening operator L, which is the way of indicating that a sparse stochastic process is a filtered version of a non-Gaussian white noise.

The delicate aspect with this simple operational description is that w is a highly singular entity that does not admit an interpretation as a conventional function of the time variable t (think of w as the stochastic counterpart of the Dirac distribution δ whose explicit definition as a tempered distribution is $\langle \delta, \varphi \rangle = \varphi(0)$ for all "test" functions φ). The mathematical framework for the correct interpretation of (11) is Gelfand's theory of generalized stochastic processes [28], which is briefly summarized as follows:

A generalized white noise is a probability measure on the dual space of a set of test functions that has the following properties:

- For a given test function φ, the statistics of the random variable $\langle w, \varphi \rangle$ do not change upon shifting φ, where w denotes a generic random element in the dual space of test functions (typically, Schwartz space of tempered distributions) and $\langle \cdot, \cdot \rangle$ denotes the inner product.
- If the test functions in the collection $\{\varphi_\beta\}_{\beta \in B}$ (B is an index set) have disjoint supports, then the random variables in $\{\langle w, \varphi_\beta \rangle\}_{\beta \in B}$ are independent.

Under some mild regularity conditions, there is a one-to-one correspondence between the infinitely divisible random variables and the white noises specified above. Thus, specifying a white noise is equivalent to prescribing the probability law of the random variable $\langle w, \varphi \rangle$ for any test function φ.

Correspondingly, we have

$$\langle s, \varphi \rangle = \langle \mathrm{L}^{-1} w, \varphi \rangle = \langle w, \mathrm{L}^{-1*} \varphi \rangle \tag{12}$$

where L^{-1*} is the adjoint operator of L^{-1}. It means that one can readily deduce the statistical distribution of $\langle s, \varphi \rangle$ given the process w.

Now, if w is SαS white noise, then the random variable $\langle w, \varphi \rangle$ has an SαS distribution whose characteristic function is given by

$$\hat{p}_{\langle w, \varphi \rangle}(\omega) = \mathbb{E}\left\{ \mathrm{e}^{\mathrm{j}\omega \langle w, \varphi \rangle} \right\} = \mathrm{e}^{-|\|\varphi\|_\alpha \omega|^\alpha} \tag{13}$$

where $\|\varphi\|_\alpha = \left(\int_{\mathbb{R}} |\varphi(t)|^\alpha dt \right)^{1/\alpha}$ is the L_α-norm of φ. In the case of an AR(1) process, we have that

$$\mathrm{L} = \mathrm{D} + \kappa \mathrm{I} \tag{14}$$

where D and I are respectively the differentiator and the identity operator; then, s in (11) is a continuous-domain SαS AR(1) process. It follows from the theory of linear systems that the impulse response of L^{-1} is the causal exponential

$$\rho_\kappa(t) = \mathrm{e}^{-\kappa t} \mathbf{1}_+(t) \tag{15}$$

where $\mathbf{1}_+(t)$ is the unit step. Thus, as a function of t, we can write

$$s(t) = (\rho_\kappa * w)(t) \tag{16}$$

where $*$ denotes the continuous-domain convolution operation. The AR(1) process is well-defined for $\kappa > 0$. The limit case $\kappa = 0$ can also be handled by setting the boundary condition $s(0) = 0$, which results in a Lévy process that is non-stationary. Realizations of AR(1) processes for $\kappa = 0.05$ and for different values of α are depicted in Fig. 1. When α decreases, the process becomes sparser in the sense that its innovation becomes more and more heavy-tailed.

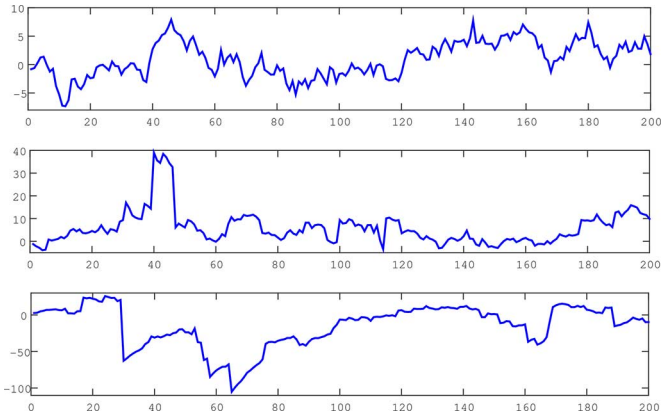

Fig. 1. Examples of AR(1) processes for different α.

B. Discretization of AR(1) Processes

Now, for a given integer k and time period T, we set

$$\varphi_k(t) = \delta(t-kT) - \mathrm{e}^{-\kappa T}\delta(t-(k-1)T) \tag{17}$$

where δ is the Dirac impulse, and define w_k as

$$w_k = \langle s, \varphi_k(t)\rangle = s(kT) - \mathrm{e}^{-\kappa T}s((k-1)T). \tag{18}$$

This means that the sampled version $\{s_k = s((k-1)T)\}_{k\in\mathbb{Z}}$ of $s(t)$ satisfies the first-order difference equation

$$s_k = \mathrm{e}^{-\kappa T}s_{k-1} + w_k. \tag{19}$$

Also, we have that

$$w_k = \langle s, \varphi_k(t)\rangle = \langle w, (\check{\rho}_\kappa * \varphi_k)(t)\rangle \tag{20}$$

where $\check{\rho}_\kappa(t) = \rho_\kappa(-t)$ is the impulse response of L^{-1*} in (12). Also,

$$(\check{\rho}_\kappa * \varphi_k)(t) = \beta_{\kappa,T}(t-kT) = \mathbf{1}_{[kT,(k+1)T)}\mathrm{e}^{-\kappa(t-kT)} \tag{21}$$

where $\mathbf{1}_{[kT,(k+1)T)}$ is the indicator function of the set $[kT,(k+1)T)$, is the exponential B-spline with parameters κ and T [14]. The fundamental property here is that the kernels $\{\beta_{\kappa,T}(\cdot - kT)\}_{k\in\mathbb{Z}}$ are shifted replicates of each other and have compact and disjoint supports. Thus, according to the definition of a white noise, $\{w_k\}_{k\in\mathbb{Z}}$ is an iid sequence of SαS random variables with the common characteristic function

$$\hat{p}_w(\omega) = \mathbb{E}\left\{\mathrm{e}^{\mathrm{j}\omega\langle w,\beta_{\kappa,T}\rangle}\right\} = \mathrm{e}^{-|\|\beta_{\kappa,T}\|_\alpha\omega|^\alpha}. \tag{22}$$

The conclusion is that a continuous-domain AR(1) process maps into the discrete AR(1) process $\{s_k\}_{k\in\mathbb{Z}}$ that is uniquely specified by (19) and (22).

We now consider N consecutive samples of the process and define the random vectors $\mathbf{s} = [s_1\cdots s_N]^\top$ and $\mathbf{w} = [w_1\cdots w_N]^\top$. This allows us to rewrite (19) as

$$\mathbf{s} = \mathbf{L}^{-1}\mathbf{w} \tag{23}$$

where $\mathbf{L}^{-1} = [\bar{l}_{ij}]_{N\times N}$ and

$$\bar{l}_{ij} = \mathrm{e}^{-\kappa T(j-i)}\cdot\mathbf{1}_{\{j\geq i\}} \tag{24}$$

which is the discrete-domain counterpart of (15).

In the next sections, we are going to study linear transforms applied to the signal s (or $\mathbf{s}$). Here, we recall a fundamental property of stable distributions that we shall use in our derivations.

Property 1 (Linear Combination of SαS *Random Variables):* Let $\bar{r} = \sum_{m=1}^M a_m r_m$ where r_m are iid SαS random variables with dispersion parameter c. Then, $\bar{r}$ is an SαS as well with dispersion parameter $\|[a_1,\ldots,a_M]\|_\alpha^\alpha c$ [17].

To establish this property, we consider M iid SαS random variables $r_1,\ldots,r_M$ with common characteristic function $\mathrm{e}^{-c|\omega|^\alpha}$, and a corresponding sequence of real-valued weights $a_1,\ldots,a_M$. Then, the characteristic function of the random variable $r^* = \sum_{m=1}^M a_m r_m$ is given by

$$\hat{p}_{r^*}(\omega) = \prod_{m=1}^M \mathrm{e}^{-c|a_m\omega|^\alpha} = \mathrm{e}^{-c\left|\left(\sum_{m=1}^M |a_m|^\alpha\right)^{1/\alpha}\omega\right|^\alpha}. \tag{25}$$

Thus, r^*, which is a linear combination of iid SαS random variables, is an SαS random variable with the same distribution as one of them multiplied by the factor $(\sum_{m=1}^M |a_m|^\alpha)^{1/\alpha}$; i.e.,

$$r^* \overset{d}{=} \left(\sum_{m=1}^M |a_m|^\alpha\right)^{1/\alpha} r_1. \tag{26}$$

C. Operator-Like Wavelets

Conventional wavelet bases act as smoothed versions of the derivative operator. To decouple the AR(1) process in (16) by a wavelet-like transform, we need to choose basis functions that essentially behave like the whitening operator L in (10). Such wavelet-like basis functions are called operator-like wavelets and can be tailored to any given differential operator L [24]. The operator-like wavelet at scale i and location k is given by

$$\psi_{i,k} = \mathrm{L}^*\phi_i(\cdot - 2^i kT), \tag{27}$$

where ϕ_i is a scale-dependent smoothing kernel and the dot is the placeholder of the index variable of the function to which the operator L^* is applied. Since $\{\psi_{i,k}\}$ is an orthonormal basis and $s = \mathrm{L}^{-1}w$, the wavelet coefficients of the signal s are

$$\begin{aligned} v_{i,k} &= \langle s, \psi_{i,k}\rangle = \langle \mathrm{L}^{-1}w, \psi_{i,k}\rangle \\ &= \langle w, \mathrm{L}^{-1*}\mathrm{L}^*\phi_i(\cdot - 2^i kT)\rangle = \langle w, \phi_i(\cdot - 2^i kT)\rangle. \end{aligned} \tag{28}$$

Based on this equality, we understand that, for any given i and for all k, the $v_{i,k}$ follows an SαS distribution with dispersion parameter $\|\phi_i\|_\alpha^\alpha$ [13]. Also, since w is independent at every point, intuitively, the level of decoupling has a direct relation to the overlap of the smoothing kernels $\phi_i(\cdot - 2^i kT)$. The operator-like wavelets proposed in [24] are very similar to Haar wavelets, except that they are piecewise exponential instead of piecewise constant (for $\kappa = 0$). Indeed, we have

$$\begin{aligned} \psi_{i,k}(t) &\propto \mathrm{e}^{-2^{-i}\kappa T}\beta_{\kappa,2^{-i}T}(t-k2^{-i}T) \\ &\quad - \beta_{\kappa,2^{-i}T}(t-(k+1)2^{-i}T) \\ &= \begin{cases} 0 & t < k2^{-i}T \\ \mathrm{e}^{-\kappa(t-(k-1)2^{-i}T)} & k2^{-i}T \leq t < (k+1)2^{-i}T \\ -\mathrm{e}^{-\kappa(t-(k+1)2^{-i}T)} & (k+1)2^{-i}T \leq t < (k+2)2^{-i}T \\ 0 & (k+2)2^{-i}T \leq t. \end{cases} \end{aligned} \tag{29}$$

For these wavelets, the supports of $\phi_{i,k}$ do not overlap within the given scale i. Thus, the wavelet coefficients at scale i are inde-

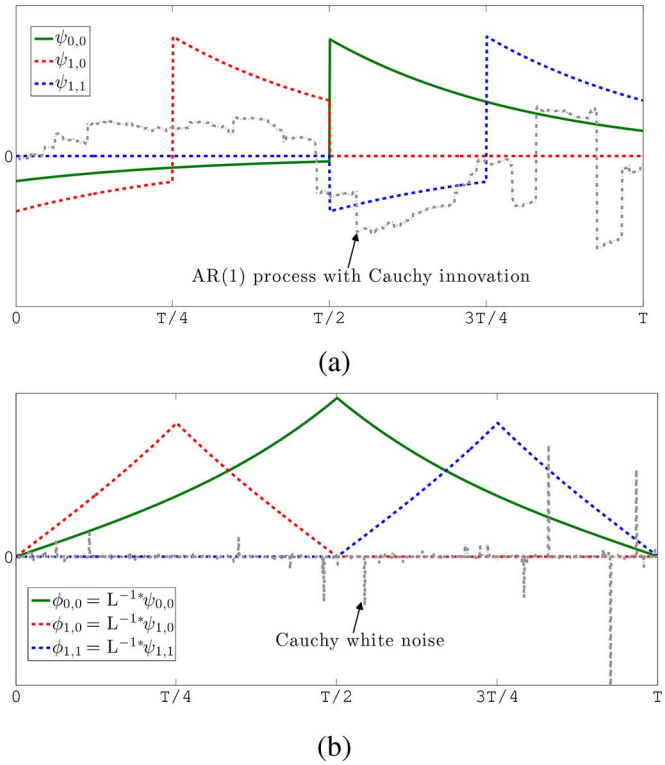

Fig. 2. Two equivalent interpretations of the wavelet analysis of a sparse process. (a) Operator-like wavelets at two consecutive scales acting on an Cauchy AR(1) process. (b) The equivalent windows (smoothing kernels) acting on the underlying Cauchy white noise. Note that $\psi_{1,0}$ and $\psi_{1,1}$ ($\phi_{1,0}$ and $\phi_{1,1}$, respectively) are non-overlapping.

pendent and identically distributed. This property suggests that this type of transform is an excellent candidate for decoupling AR(1) processes. The illustration of plugging these wavelets into (28) is given in Fig. 2.

IV. Search for the Optimal Transformation

From now on, we assume that the signal vector $\mathbf{s} = [s_1 \cdots s_N]^\top$ with $s_k = s((k-1)T)$ is obtained from the samples of an SαS AR(1) process and satisfies the discrete innovation model (19). The representation of the signal $\mathbf{s}$ in (23) in the transform domain is denoted by $\mathbf{y} = [y_1 \cdots y_N]^\top = \mathbf{Hs}$, where $\mathbf{H} = [h_{ij}]_{N\times N}$ is the underlying orthogonal transformation matrix (e.g., DCT, wavelet transform). The idea is now to rely on Property 1 to derive the explicit form of the proposed performance criteria under the SαS hypothesis. This, in turn, will allow us to determine the optimal transform (ICA solution) based on numerical optimization.

A. Optimizing Coding Performance

Let us now use (3) to characterize the performance of a given transformation matrix $\mathbf{H}$. First, we simplify (3) to

$$\begin{aligned}\mathrm{R}(\mathbf{H}) &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{H}(y_n) - \frac{1}{N}\mathbb{H}(\mathbf{y}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{H}(y_n) - \mathbb{H}(w_1) - \frac{1}{N}\log\det\mathbf{HL}^{-1}, \quad (30)\end{aligned}$$

where $\mathbb{H}(\cdot)$ is the differential entropy defined in (1). Also, we observe that $\log\det\mathbf{HL}^{-1} = 0$. In addition, since the w_m is α-stable, according to Property 1, we can write

$$y_n \overset{d}{=} \bar{h}_n w_1, \quad (31)$$

where $\bar{h}_n$ is the α-(pseudo)norm of the nth row of $\mathbf{HL}^{-1}$ given by

$$\bar{h}_n = \left(\sum_{r=1}^{N}\left|\sum_{m=1}^{N} h_{nm}\bar{l}_{mr}\right|^{\alpha}\right)^{\frac{1}{\alpha}}. \quad (32)$$

It follows that

$$\mathrm{R}(\mathbf{H}) = \frac{1}{N}\sum_{n=1}^{N}\log\bar{h}_n, \quad (33)$$

which can be readily calculated for any given $\mathbf{H}$.

Note 1: This criterion is reminiscent of the sum-of-dispersion criterion $\sum_{n=1}^{N}\bar{h}_n$ which is frequently used in the study of α-stable stochastic processes [29], [30]. However, unlike (33), the latter dispersion criterion does not have a direct information-theoretic interpretation.

B. Optimizing Denoising Performance

As a second option, we use the criterion (9) to measure the performance of a given transform matrix $\mathbf{H}$ for the denoising task. Similar to the case in (30), it can be simplified to

$$\mathrm{MSE}(\mathbf{H}) = \sigma^2 - \frac{\sigma^4}{N}\sum_{n=1}^{N}\int\frac{\left(p'_{\tilde{y}_n}(\tilde{y}_n)\right)^2}{p_{\tilde{y}_n}(\tilde{y}_n)}\mathrm{d}\tilde{y}_n, \quad (34)$$

in which σ^2 is the noise variance and $\tilde{y}_n$ is the nth entry of

$$\tilde{\mathbf{y}} = \mathbf{Hz} = \mathbf{Hs} + \mathbf{Hn} = \mathbf{y} + \tilde{\mathbf{n}}. \quad (35)$$

Since $\mathbf{H}$ is a unitary matrix, $\tilde{\mathbf{n}}$ has the same distribution as $\mathbf{n}$. Also, according to (31),

$$\tilde{y}_n \overset{d}{=} \bar{h}_n w_1 + n_1, \quad (36)$$

where n_1 is a standard Gaussian random variable. This allows us to deduce the pdf expression

$$p_{\tilde{y}_n}(y) = \frac{1}{\bar{h}_n}p_{w_1}\left(\frac{y}{\bar{h}_n}\right) * p_{n_1}(y) \quad (37)$$

which involves the convolution of a rescaled SαS law with a Gaussian of standard deviation σ. Thus, (34) is calculable through one-dimensional integrals.

C. Numerical Implementation

Based on (33) and (34), we can now attempt to find the optimal transformation $\mathbf{H}_{\mathrm{ICA}}$ by minimizing these expressions over the space of all orthonormal matrices of size N.

To guide this optimization process, we first derive the gradient of the cost functions R and MSE with respect to $\mathbf{H}$. Specifically, according to (32) and (33), the partial derivative of $\mathrm{R}(\mathbf{H})$ is

$$\frac{\partial\mathrm{R}}{\partial h_{ij}} = \frac{1}{N\alpha\bar{h}_i^{\alpha}}\frac{\partial\bar{h}_i^{\alpha}}{\partial h_{ij}} \quad (38)$$

where

$$\frac{\partial \bar{h}_i^{\alpha}}{\partial h_{ij}} = \alpha \sum_{r=1}^{N} l_{jr} \operatorname{sgn}\left(\sum_{n=1}^{N} h_{ik} l_{kr}\right) \left|\sum_{n=1}^{N} h_{ik} l_{kr}\right|^{\alpha-1}. \quad (39)$$

Also, the partial derivative of $\mathrm{MSE}(\mathbf{H})$ in (34) is

$$\begin{aligned}\frac{\partial \mathrm{MSE}}{\partial h_{ij}} &= -\frac{\sigma^4}{N} \frac{\partial}{\partial \bar{h}_i} \int \frac{\left(p_{\tilde{y}_i}^{(1)}(u)\right)^2}{p_{\tilde{y}_i}(u)} \mathrm{d}u \times \frac{\bar{h}_i^{1-\alpha}}{\alpha} \frac{\partial \bar{h}_i^{\alpha}}{\partial h_{ij}} \\ &= -\frac{\sigma^4}{N} \left(2 \int \frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}^{(1)}(u) \frac{p_{\tilde{y}_i}^{(1)}(u)}{p_{\tilde{y}_i}(u)} \mathrm{d}u \right. \\ &\quad \left. - \int \frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}(u) \left(\frac{p_{\tilde{y}_i}^{(1)}(u)}{p_{\tilde{y}_i}(u)}\right)^2 \mathrm{d}u \right) \frac{\bar{h}_i^{1-\alpha}}{\alpha} \frac{\partial \bar{h}_i^{\alpha}}{\partial h_{ij}} \quad (40)\end{aligned}$$

in which $p_{\tilde{y}_i}^{(k)}(y)$ is the kth derivative of $p_{\tilde{y}_i}(y)$ which, according to (37), can be written as

$$p_{\tilde{y}_i}^{(k)}(y) = p_{y_i}(y) * \frac{\mathrm{d}^k}{\mathrm{d}y^k} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{y^2}{2\sigma^2}} \right). \quad (41)$$

Also, we have that

$$\frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}(y) = -\frac{1}{\bar{h}_i} p_{\tilde{y}_i}(y) - \frac{y}{\bar{h}_i} p_{\tilde{y}_i}^{(1)}(y) - \frac{1}{\bar{h}_i} p_{\tilde{y}_i}^{(2)}(y) \quad (42)$$

and

$$\frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}^{(1)}(y) = -\frac{2}{\bar{h}_i} p_{\tilde{y}_i}^{(1)}(y) - \frac{y}{\bar{h}_i} p_{\tilde{y}_i}^{(2)}(y) - \frac{1}{\bar{h}_i} p_{\tilde{y}_i}^{(3)}(y). \quad (43)$$

Now, since the y_i have nice characteristic functions, we can calculate (41) efficiently through the inverse Fourier transform

$$p_{y_i}^{(k)}(y) = \mathcal{F}_{\omega}^{-1} \left\{ (\mathrm{j}\omega)^k \mathrm{e}^{-|\bar{h}_i \omega|^{\alpha} - \frac{\sigma^2}{2}\omega^2} \right\}(y) \quad (44)$$

using the FFT algorithm.

Thus, we can use gradient-based optimization to obtain the optimal transformations for different values of κ, α, and N. For our experiments, we implemented a gradient-descent algorithm with adaptive step size to efficiently find the optimal transform matrix. Since the transform matrix may deviate from the space of unitary matrices, after each step, we project it on that space using the method explained in Appendix A. Given the measure of independence C (i.e., R or MSE), the algorithm is as follows:

Algorithm 1: Steepest-Descent Algorithm with Adaptive Step-Size to Apply ICA to Discrete SαS AR(1) Processes

input: N, α, κ
initialize: $\mathbf{H}_{\mathrm{old}}$, μ, $a \in [1, +\infty)$ and $b \in [0, 1]$
repeat
  $\widetilde{\mathbf{H}}_{\mathrm{new}} = \mathbf{H}_{\mathrm{old}} - \mu \nabla C|_{\mathbf{H}_{\mathrm{old}}}$
  Set $\mathbf{H}_{\mathrm{new}}$ to the projection of $\widetilde{\mathbf{H}}_{\mathrm{new}}$ onto the space of unitary matrices
  if $C(\mathbf{H}_{\mathrm{new}}) < C(\mathbf{H}_{\mathrm{old}})$ **then**
    $\mathbf{H}_{\mathrm{old}} \leftarrow \mathbf{H}_{\mathrm{new}}$
    $\mu \leftarrow a \cdot \mu$
  else
    $\mathbf{H}_{\mathrm{new}} \leftarrow \mathbf{H}_{\mathrm{old}}$
    $\mu \leftarrow b \cdot \mu$
  end if
until convergence
return $\mathbf{H}_{\mathrm{new}}$

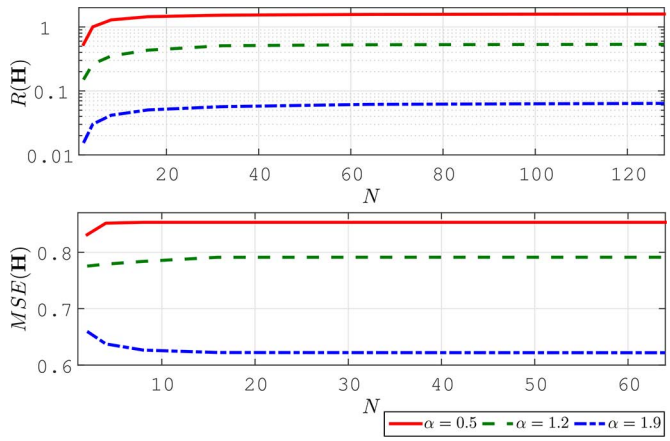

Fig. 3. Minimum value of $\mathrm{R}(\mathbf{H})$ and $\mathrm{MSE}(\mathbf{H})$ for Lévy processes as a function of N for different values of α. In the second plot $\sigma^2 = 1$.

Algorithm 1 can be viewed as a model-based version of ICA. We take advantage of the underlying stochastic model to derive an optimal solution based on the minimization of (33) and (34), which involves the computation of ℓ_α-norms of the transformation matrix. By contrast, the classical version of ICA is usually determined empirically based on the observations of a process, but the ultimate aim is similar; namely, the decoupling of the data vector.

V. Results for Different Transformations

The majority of experiments on ICA published in the literature are data-driven. The present formulation, by contrast, is model-based so that it does not require the generation of signal samples. To make an analogy, it is to ICA what the Karhunen-Loève transform is to principal components (PCA). We can therefore rely on (33)–(34) to compute the performance of a transform analytically. Also, the optimal transform (referred to as ICA) is found numerically by running Algorithm 1. We recall that our theoretical figures of merit are relevant to practical signal processing: the first (mutual information) gives in a direct measure of the coding gain in a compression experiment, while the second measures the signal-to-noise ratio (SNR) improvement for signal denoising, as justified in Section II.

Initially, we investigate the effect of the signal length N on the value of R and MSE. We consider the case of a Lévy process (i.e., $\kappa = 0$) and numerically optimize the criteria for different α and plot it as a function of N. Results are depicted in Fig. 3. As we see, the criteria values converge quickly to their asymptotic values. Thus, for the remainder of the experiments, we choose $N = 64$. This is a block size that is reasonable computationally and large enough to be representative of the asymptotic regime.

Then, we investigate the performance of different transforms for various processes. First, we focus on the Lévy processes. In this case, the operator-like wavelet transform is the classical Haar wavelet transform (HWT). The performance criteria R and MSE as a function of α for various transforms are plotted in Figs. 4 and 5, respectively. The considered transformations are as follows: identity as the baseline, discrete cosine transform (DCT), Haar wavelet transform (HWT), and optimal solution (ICA) provided by the proposed algorithm. In the case of $\alpha = 2$ (Gaussian scenario), the process s is a Brownian motion whose KLT is a sinusoidal transform that is known ana-

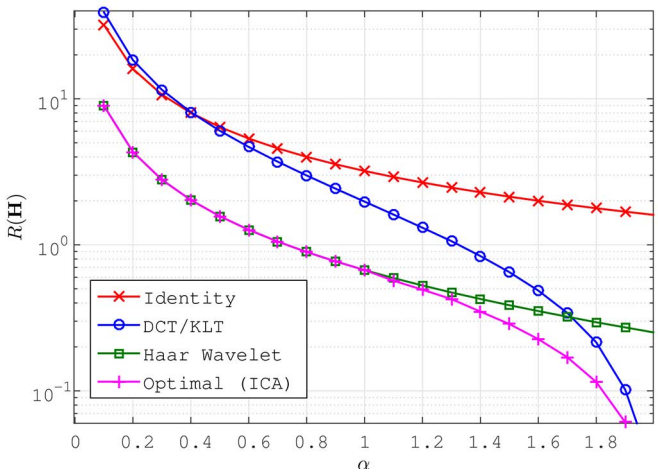

Fig. 4. $\mathrm{R}(\mathbf{H})$ of Lévy processes versus α when $N = 64$ for different $\mathbf{H}$.

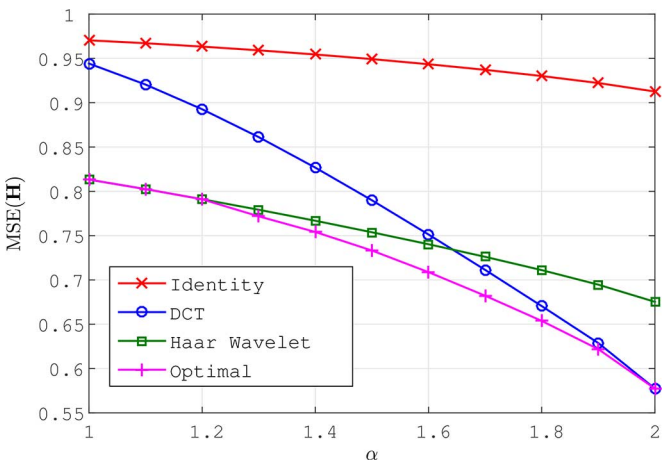

Fig. 5. $\mathrm{MSE}(\mathbf{H})$ of Lévy processes versus α when $N = 64$ for different $\mathbf{H}$ when $\sigma^2 = 1$.

lytically [31]. In this case, the DCT and the optimal transform converge to the KLT since being decorrelated is equivalent to being independent. We see this coincidence in both Figs. 4 and 5. The vanishing of R at $\alpha = 2$ indicates perfect decoupling. By contrast, as α decreases, neither the DCT nor the optimal transform decouples the signal completely. The latter means that there is no unitary transform that completely decouples stable non-Gaussian Lévy processes. However, we see that, based on both criteria R and MSE, and as α decreases, the DCT becomes less favorable while the performance of the HWT gets closer to the optimal one. Moreover, Figs. 4 and 5 even suggest that the Haar wavelet transform is equivalent to the ICA solution for $\alpha \leq 1$.

Also, to see the transition from sinusoidal bases to Haar wavelet bases, we plot the optimal basis which is obtained by the proposed algorithm at two consequent scales. In Fig. 6, we see the progressive evolution of the ICA solution from the sinusoidal basis to the Haar basis while changing the parameter α of the model.

Next, we consider a stationary AR(1) process with $\mathrm{e}^{-\kappa T} = 0.9$ and $n = 64$. For $\alpha = 2$, we get the well-known classical Gaussian AR(1) process for which the DCT is known to be asymptotically optimal [1], [3]. For such a process, the operator-like wavelet is known before hand and given by (29). The performance criterion R versus α for the DCT, the HWT, the operator-like wavelet matched to the process, and the optimal ICA solution are plotted in Fig. 7. Here too we see that, for $\alpha = 2$, ICA is equivalent the DCT. But, as α decreases, the DCT loses its optimality and the matched operator-like wavelet becomes closer to optimum. Again, we observe that, for $\alpha \leq 1$, the ICA solution is the matched operator-like wavelet described in Section III.C. The fact that the matched operator-like wavelet outperforms the HWT shows the benefit of the tuning of the wavelet to the differential characteristics of the process. Also, as shown in Fig. 8, experimentally determined ICA basis functions for $\alpha = 1$ are indistinguishable from the wavelets in Fig. 2.

To substantiate those findings, we present a theorem that states that, based on the above mentioned criteria and for any $\alpha < 2$, the operator-like wavelet transform outperforms the DCT (or, equivalently, the KLT associated with the Gaussian member of the family) as the block-size N tends to infinity.

Theorem 1: If $\alpha < 2$ and $\kappa \geq 0$, we have that

$$\lim_{N\to\infty} \mathrm{R}(\mathrm{OpWT}) < \lim_{N\to\infty} \mathrm{R}(\mathrm{DCT}) = \infty \tag{45}$$

and

$$\lim_{N\to\infty} \mathrm{MSE}(\mathrm{OpWT}) < \lim_{N\to\infty} \mathrm{MSE}(\mathrm{DCT}) = \sigma^2, \tag{46}$$

where OpWT stands for the operator-like wavelet transform. The proof is given in Appendix B.

In addition, this theorem states that, for $\alpha < 2$ and as N tends to ∞, the performance of the DCT is equivalent to the trivial identity operator. This is surprising because, since the DCT is optimal for the Gaussian case $(\alpha = 2)$, one may expect that it has a good result for other AR(1) processes. However, although this theorem does not assert that operator-like wavelets are the optimal basis, it still shows that, by applying them, we obtain better performance than trivial transformations. Also, through simulations, we observed that operator-like wavelets are close to optimal transform as α gets smaller. In such extreme scenarios, the probabilities densities of the signal and of its transformed-domain coefficients are extremely heavy-tailed which conforms with a statistical notion of sparsity [15], [16].

It is worth mentioning that, in addition to the gain in performance, operator-like wavelets are cheaper to compute than the DCT. They can be implemented with the same type of filter-bank algorithm as the Haar transform, the only difference being that the filters are scale-dependent. The resulting cost is of $\mathcal{O}(N)$ (two operations per coefficient) which compares favorably with the $\mathcal{O}(N \log N)$ of the DCT. Using operator-like wavelets is also immensely more efficient than deploying the full ICA machinery. The latter requires the estimation of the transform and then its full matrix computation $(\mathcal{O}(N^2))$ which cannot benefit from any acceleration due to lack of structure.

VI. Summary and Future Studies

In this paper, we focused on the simplest version (first-order differential system with an SαS excitation) of the sparse stochastic processes which have been proposed by Unser *et al.* [13], [14]. Because of the underlying innovation model and the properties of SαS random variables, we could obtain a closed-form formula for the performance of different transform-domain

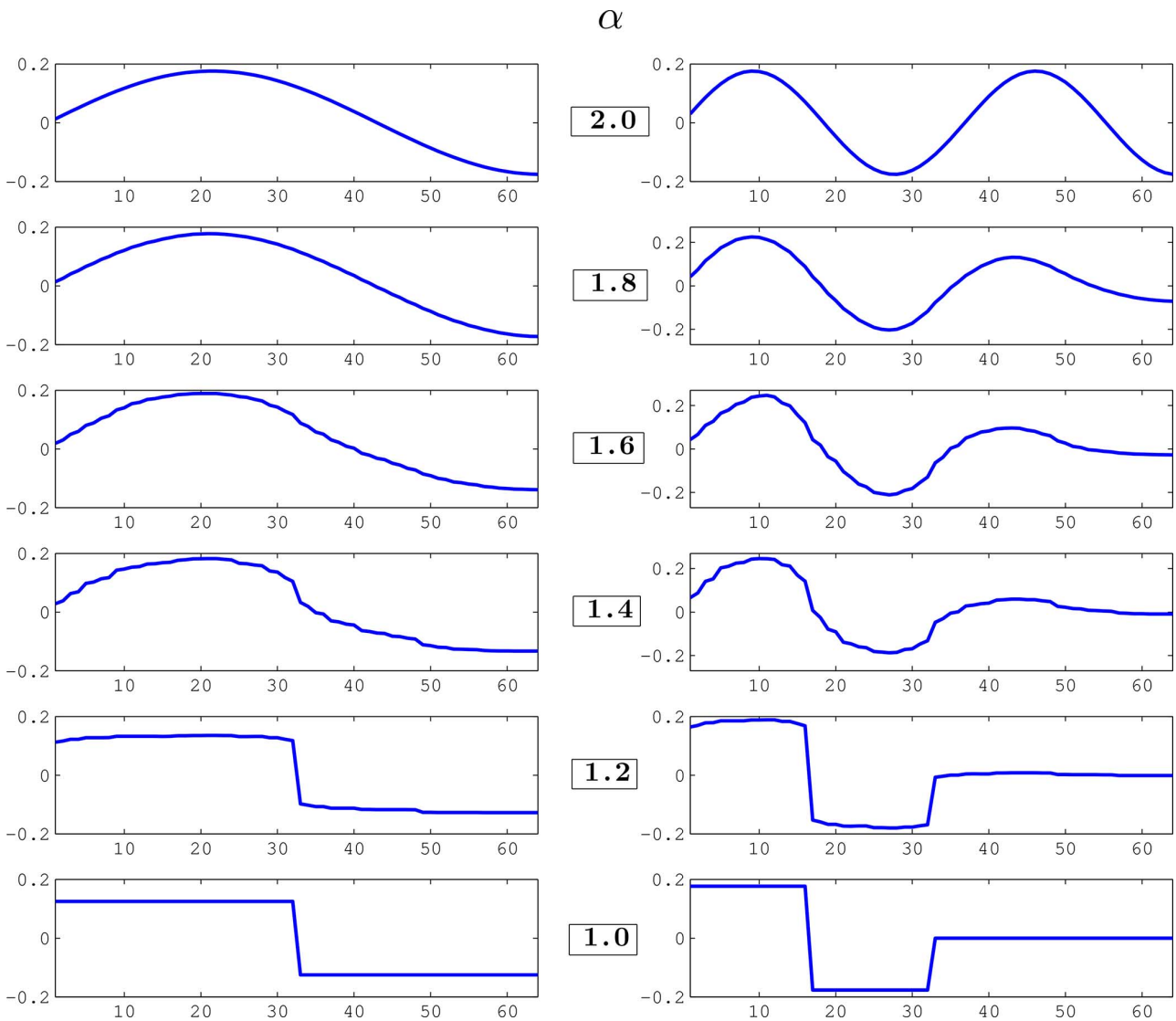

Fig. 6. Two rows of the optimal $\mathbf{H}$ (ICA) for $\alpha = 2$ down to 1 when $N = 64$. In each row, we see the evolution from sinusoidal waves to Haar wavelets by increasing the sparsity of the underlying innovation process.

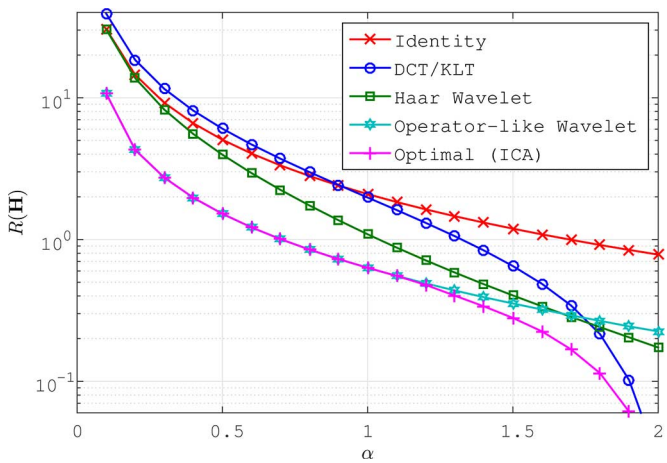

Fig. 7. $\mathrm{R}(\mathbf{H})$ versus α when $\mathrm{e}^{-\kappa T} = 0.9$ and $N = 64$ for different $\mathbf{H}$.

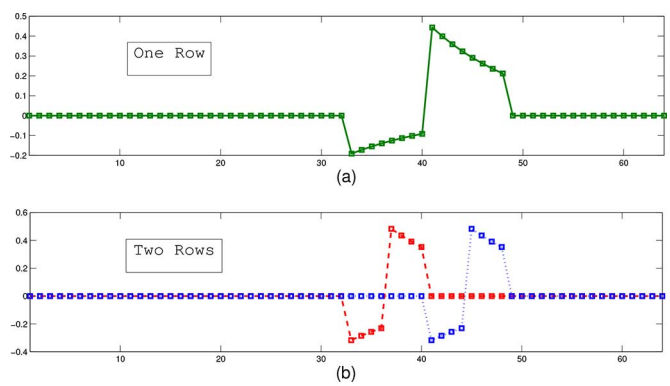

Fig. 8. Three rows of the optimal $\mathbf{H}$ for $\alpha = 1$ and $N = 64$. Parts (a) and (b) show the dyadic structure of the wavelets.

representations and characterize the optimal transform. This is a novel model-based point of view for ICA. We proved that operator-like wavelets are better than sinusoidal transforms for decoupling the sparse AR(1) processes ($\alpha < 2$). This result is remarkable since sinusoidal bases are known to be asymptotically optimal for the classical case of $\alpha = 2$ [1], [3]. Moreover, we showed that, for very sparse excitations ($\alpha \lesssim 1$), operator-like wavelets are equivalent to the ICA. As far as we know, this is the

first theoretical results on the optimality of wavelet-like bases for a given class of stochastic processes.

Another interesting aspect of this study is that it gives a unified framework for Fourier-type transforms and a class of wavelet transforms. Now, the Fourier transform and the wavelet transforms were based on two different intuitions and philosophies. However, here we have a model in which we obtain both transform families just by changing the underlying parameters.

The next step in this line of research is to investigate the extent to which these findings can be generalized to other white noises or higher-order differential operators. Also, studying the problem in the original continuous domain would be theoretically very valuable.

Appendix A
Projection on the Space of Unitary Matrices

Suppose that $\mathbf{A}$ is an $N \times N$ matrix. Our goal is to find the unitary matrix $\mathbf{H}^*$ that is the closest to $\mathbf{A}$ in Frobenius norm, in the sense that

$$\mathbf{H}^* = \arg\min_{\mathbf{H}} \|\mathbf{A} - \mathbf{H}\|_F. \tag{47}$$

According to singular-value decomposition (SVD), we can write $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Lambda}$ is a diagonal matrix with nonnegative diagonal entries.

Since the Frobenius norm is unitarily invariant, we have that

$$\|\mathbf{A} - \mathbf{H}\|_F = \|\mathbf{\Lambda} - \mathbf{U}^\top\mathbf{H}\mathbf{V}\|_F \tag{48}$$

in which $\mathbf{U}^\top\mathbf{H}\mathbf{V}$ is a unitary matrix that we call $\mathbf{K}$. The expansion of the right-hand side of (48) gives

$$\begin{aligned}\|\mathbf{\Lambda} - \mathbf{K}\|_F^2 &= \sum_{1\le i,j\le N} k_{ij}^2 + \sum_{i=1}^N \lambda_{ii}^2 - 2\sum_{i=1}^N \lambda_{ii}k_{ii} \\ &= N + \sum_{i=1}^N \lambda_{ii}^2 - 2\sum_{i=1}^N \lambda_{ii}k_{ii}. \end{aligned}\tag{49}$$

Since $\mathbf{K}$ is unitary, $|k_{ii}| \le 1$ for $i = 1, \dots, N$. Thus, setting $k_{ii} = 1$, which means setting $\mathbf{K} = \mathbf{I}$, minimizes (49). Consequently, the projection of $\mathbf{A}$ on the space of unitary matrices is $\mathbf{H}^* = \mathbf{U}\mathbf{V}^\top$.

Appendix B
Proof of Theorem 1

A. Proof of Part 1 ((45))

According to (33), we have that

$$\begin{aligned}\mathrm{R}(\mathbf{H}) &= \frac{1}{N}\sum_{n=1}^N \log \bar{h}_n = \frac{1}{N}\sum_{n=1}^N \log\left(\frac{1}{\bar{h}_n^{-1}}\right) \\ &= \int_{\mathbb{R}} \log\left(\frac{1}{\gamma}\right) p(\gamma)\mathrm{d}\gamma \end{aligned}\tag{50}$$

in which $p(\cdot)$ is the empirical distribution of $\bar{h}_n^{-1}$.

According to SVD, we can write $\mathbf{L}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ where $\mathbf{\Lambda}$ is a diagonal matrix with λ_i as diagonal entries. Taking $\mathbf{s}$ in the KLT domain is equivalent to multiplying it by $\mathbf{U}^\top$. The eigenvalues of the covariance of AR(1) matrices are known in closed form and are given by [32] and [33], for $\kappa \ge 0$, as

$$|\lambda_i|^{-1} = \sqrt{(1-\mathrm{e}^{-\kappa T})^2 + 4\mathrm{e}^{-\kappa T}\sin^2\left(\frac{\omega_i}{2}\right)} \tag{51}$$

and

$$\begin{aligned} v_{ij} = &\sqrt{\frac{2}{N + (1-\mathrm{e}^{-2\kappa T})\lambda_i^2}} \\ &\times \sin\left(\omega_i\left(j - \frac{N+1}{2}\right) + i\frac{\pi}{2}\right) \end{aligned}\tag{52}$$

in which ω_i, $i = 1, \dots, N$, is the ith positive root of

$$\tan(N\omega) = -\frac{(1-\mathrm{e}^{-2\kappa T})\sin\omega}{\cos\omega - 2\mathrm{e}^{-\kappa T} + \mathrm{e}^{-2\kappa T}\cos\omega}. \tag{53}$$

Since $\tan(N\omega)$ is an injective function that sweeps the whole domain of the real numbers while $\omega \in [\frac{i-1}{N}\pi, \frac{i}{N}\pi]$, for $i = 1, \dots, N$, (53) has a single root in each of such intervals. Thus, as N tends to infinity, the empirical distribution of the ω_i tends to the uniform distribution on $[0, \pi]$. Then, starting from (51), one can obtain the limit empirical distribution of $|\lambda_i|$ as

$$p_\lambda(\lambda) = \frac{2}{\pi}\frac{\lambda}{\sqrt{\lambda^2 - (1-\mathrm{e}^{-\kappa T})^2}\sqrt{(1+\mathrm{e}^{-\kappa T})^2 - \lambda^2}}. \tag{54}$$

Now, $\sum_{j=1}^N v_{ij}^2 = 1$ means that

$$\sum_{j=1}^N \left|\sin\left(\omega_i\left(j - \frac{N+1}{2}\right) + i\frac{\pi}{2}\right)\right|^2 \sim \mathcal{O}(N) \tag{55}$$

as N tends to infinity. But, for $\alpha < 2$, we have that

$$\begin{aligned}&\left(\sum_{j=1}^N \left|\sin\left(\omega_i\left(j - \frac{N+1}{2}\right) + i\frac{\pi}{2}\right)\right|^\alpha\right)^{\frac{1}{\alpha}} \\ &\ge \left(\sum_{j=1}^N \left|\sin\left(\omega_i\left(j - \frac{N+1}{2}\right) + i\frac{\pi}{2}\right)\right|^2\right)^{\frac{1}{\alpha}} \sim \mathcal{O}\left(N^{\frac{1}{\alpha}}\right). \end{aligned}\tag{56}$$

Thus, for $\alpha < 2$, $(\sum_{j=1}^N |v_{ij}|^\alpha)^{\frac{1}{\alpha}}$ grows faster than $\mathcal{O}(N^{\frac{1}{\alpha}-\frac{1}{2}})$ and thus tends to infinity as N tends to infinity. Consequently, the limit empirical distribution of $\bar{h}_i^{-1}$ can be represented as

$$p(\gamma) = \begin{cases} \frac{2}{\pi}\frac{\gamma}{\sqrt{\gamma^2 - (1-\mathrm{e}^{-\kappa T})^2}\sqrt{(1+\mathrm{e}^{-\kappa T})^2 - \gamma^2}} & \alpha = 2 \\ \delta(\gamma) & \alpha \ne 2. \end{cases} \tag{57}$$

By plugging this result into (50), we conclude that, for $\alpha < 2$, $\lim_{N\to\infty} \mathrm{R}(\mathrm{KLT}) = \infty$. This completes the proof of the right-hand side.

Now, for the proof of the left-hand side, we need to specify the matrix $\mathbf{H}$ for the operator-like wavelet transform. This matrix is given by the recursive construction

$$\begin{aligned}\mathbf{H}_k = &\mathrm{diag}\left(\sqrt{\frac{1-\mathrm{e}^{-2\kappa T}}{1-\mathrm{e}^{-2^{k+1}\kappa T}}}, \sqrt{\frac{1-\mathrm{e}^{-2\kappa T}}{1-\mathrm{e}^{-2^{k+1}\kappa T}}}, \underbrace{1, \dots, 1}_{2^k-2}\right) \\ &\times \begin{bmatrix} \boldsymbol{\ell}_{k-1} & \mathrm{e}^{-2^{k-1}\kappa T}\boldsymbol{\ell}_{k-1} \\ -\mathrm{e}^{-2^{k-1}\kappa T}\boldsymbol{\ell}_{k-1} & \boldsymbol{\ell}_{k-1} \\ \mathbf{H}'_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}'_{k-1} \end{bmatrix} \end{aligned}\tag{58}$$

in which $\mathbf{H}'_{k-1}$ is the matrix $\mathbf{H}_{k-1}$ omitting the first row and $\boldsymbol{\ell}_{k-1} = [1, \mathrm{e}^{-\kappa T}, \ldots, \mathrm{e}^{-(2^{k-1}-1)\kappa T}]$. Also, $\mathbf{H}_0 = [1]$. Let us denote the empirical distribution of $\bar{h}_i^{-1}$ (the reciprocal of the α-(pseudo) norm of the rows of $\mathbf{H}_k\mathbf{L}_{2^k}$) by $p_k(\gamma) = \sum_{i=1}^{k} p_i \delta(\gamma - \gamma_i)$. Now, for the sequence of p_i and γ_i, with respect to k, we have the following recursive relation:

- Replace p_{k-1} by $\left(\frac{p_{k-1}}{2}, \frac{p_{k-1}}{2}\right)$
- Remove γ_{k-1}. Then, if $\kappa > 0$, set

$$\gamma_{k-1} = \sqrt{\frac{1-\mathrm{e}^{-2^{k+1}\kappa T}}{1-\mathrm{e}^{-2\kappa T}}} \times \left(\sum_{i=-2^{k-1}+1}^{2^{k-1}} \left(\frac{\mathrm{e}^{-|i|\kappa T} - \mathrm{e}^{-(2^k-|i|)\kappa T}}{1-\mathrm{e}^{-2\kappa T}} \right)^{\alpha} \right)^{-\frac{1}{\alpha}} \quad (59)$$

and

$$\gamma_k = \sqrt{\frac{1-\mathrm{e}^{-2^{k+1}\kappa T}}{1-\mathrm{e}^{-2\kappa T}}} \left(\sum_{i=1}^{2^k} \left(\frac{1-\mathrm{e}^{-2i\kappa T}}{1-\mathrm{e}^{-2\kappa T}} \right)^{\alpha} \right)^{-\frac{1}{\alpha}} \quad (60)$$

else, if $\kappa = 0$, set

$$\gamma_{k-1} = 2^{\frac{k}{2}} \left(\sum_{i=-2^{k-1}+1}^{2^{k-1}} \left(2^{k-1} - |i|\right)^{\alpha} \right)^{-\frac{1}{\alpha}} \quad (61)$$

and

$$\gamma_k = 2^{\frac{k}{2}} \left(\sum_{i=1}^{2^k} i^{\alpha} \right)^{-\frac{1}{\alpha}}. \quad (62)$$

Consequently, according to (50), we have that

$$\lim_{n\to\infty} \mathrm{R}(\mathrm{HWT}) = \sum_{k=1}^{\infty} 2^{-k} \log \gamma_k^{-1}. \quad (63)$$

However, for the case $\kappa > 0$ and $k < N$,

$$\begin{aligned} \gamma_k^{-1} &\le \frac{2\left(\left(2^k-1\right)\left(1-\mathrm{e}^{-2^k\kappa T}\right)^{\alpha}\right)^{\frac{1}{\alpha}}}{\sqrt{\left(1-\mathrm{e}^{-2\kappa T}\right)\left(1-\mathrm{e}^{-2^{k+1}\kappa T}\right)}} \\ &\le \frac{2}{\sqrt{1-\mathrm{e}^{-2\kappa T}}} \sqrt{\frac{1-\mathrm{e}^{-2^k\kappa T}}{1+\mathrm{e}^{-2^k\kappa T}}} (2^k-1)^{\frac{1}{\alpha}} \\ &\le \frac{2^{1+\frac{k}{\alpha}}}{\sqrt{1-\mathrm{e}^{-2\kappa T}}}. \end{aligned} \quad (64)$$

Thus,

$$\begin{aligned} \lim_{n\to\infty} \mathrm{R}(\mathrm{HWT}) &\le \sum_{k=1}^{\infty} 2^{-k} \log \frac{2^{1+\frac{k}{\alpha}}}{\sqrt{1-\mathrm{e}^{-2\kappa T}}} \\ &= \left(\frac{2}{\alpha} + \frac{1}{2} \log \frac{1}{1-\mathrm{e}^{-2\kappa T}} \right) \log 2. \end{aligned} \quad (65)$$

For the case $\kappa = 0$ and $k < N$,

$$\begin{aligned} \gamma_k^{-1} &\le 2^{-\frac{k}{2}} \left((2^k-1)(2^{k-1})^{\alpha} \right)^{\frac{1}{\alpha}} \\ &\le 2^{\frac{k}{2}+\frac{k}{\alpha}-1}. \end{aligned} \quad (66)$$

Thus,

$$\lim_{n\to\infty} \mathrm{R}(\mathrm{HWT}) \le \sum_{k=1}^{\infty} 2^{-k} \log 2^{\frac{k}{2}+\frac{k}{\alpha}-1} = \frac{2}{\alpha} \log 2. \quad (67)$$

Therefore, the proof is complete.

B. Proof of Part 2 ((46))

Proof: We have that

$$\mathrm{MSE}(\mathbf{H}) = \frac{1}{N} \sum_{n=1}^{N} \nu\left(\bar{h}_n^{-1}\right) = \int_{\mathbb{R}} \nu(\gamma^{-1}) p(\gamma) \mathrm{d}\gamma \quad (68)$$

in which $\nu(\gamma^{-1})$ is the MMSE of the estimating w from s in the scalar problem

$$s = \gamma^{-1} w + z, \quad (69)$$

where w is a stable random variable with characteristic function $\hat{p}_w(\omega) = \exp(-|\omega|^{\alpha})$ and z is a Gaussian random variable with variance σ^2. We know that $\nu(\,\cdot\,)$ is a monotone continuous function that vanishes at zero and tends to σ^2 asymptotically. Also, $p(\,\cdot\,)$ is the empirical distribution of the reciprocals of $\bar{h}_i$ in (32). The proof is then essentially the same as the one of Theorem 1 but simpler since the function $\nu(\,\cdot\,)$ is bounded.

For $\mathbf{H}$ equal to Fourier transform, the limiting $p(\gamma)$ was given in (57). Thus, for $\alpha < 2$, as n tends to infinity, $\mathrm{MSE}(\mathbf{H})$ tends to σ^2. This completes the proof of the right-hand side.

For the case that $\mathbf{H}$ is the operator-like wavelet transform, the limit is $p(\gamma) = \sum_{k=1}^{\infty} p_k \delta(\gamma - \gamma_k)$ where $p_k = 2^{-k}$ and γ_k were given in (59)–(62). Thus, we have that

$$\mathrm{MSE}(\mathrm{OpWT}) = \sum_{k=1}^{\infty} 2^{-k} \nu(\gamma_k^{-1}) \le \frac{1}{2}\nu(\gamma_1^{-1}) + \frac{\sigma^2}{2}. \quad (70)$$

But, obviously, $\gamma_1^{-1} < \infty$; hence, $\nu(\gamma_1^{-1}) < \sigma^2$, which completes the proof.

References

[1] J. Pearl, "On coding and filtering stationary signals by discrete Fourier transforms," *IEEE Trans. Inf. Theory*, vol. 19, no. 2, pp. 229–232, Mar. 1973.

[2] N. Ahmed, "Discrete cosine transform," *IEEE Trans. Commun.*, vol. 23, no. 1, pp. 90–93, Sept. 1974.

[3] M. Unser, "On the approximation of the discrete Karhunen-Loève transform for stationary processes," *Signal Process.*, vol. 7, no. 3, pp. 231–249, Dec. 1984.

[4] M. Hamidi and J. Pearl, "Comparison of the cosine and Fourier transforms of Markov-1 signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 428–429, 1976.

[5] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Norwell, MA, USA: Kluwer, 2001.

[6] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," in *Proceedings of Symposia in Applied Mathematics*. Providence, RI, USA: Amer. Math. Soc., 1993, vol. 47, pp. 173–205.

[7] C. Taswell, "The what, how, and why of wavelet shrinkage denoising," *Comput. Sci. Eng.*, vol. 2, no. 3, pp. 12–19, May/June 2000.

[8] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, Jun. 13, 1996.

[9] J. F. Cardoso and D. L. Donoho, "Some experiments on independent component analysis of non-Gaussian processes," in *Proc. IEEE Signal Process. Workshop Higher-Order Statist.*, Caesarea, Jun. 14–16, 1999, pp. 74–77.

[10] Y. Meyer, Wavelets and Applications. Lecture at CIRM Luminy Meeting, Luminy, France, Mar. 1992.

[11] P. Flandrin, "On the spectrum of fractional Brownian motions," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 197–199, Jan. 1989.
[12] R. A. Devore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, Jan. 1998.
[13] M. Unser, P. D. Tafti, and Q. Sun, "A unified formulation of Gaussian vs. sparse stochastic processes—Part I: Continuous-domain theory," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1945–1962, Mar. 2014.
[14] M. Unser, P. D. Tafti, A. Amini, and H. Kirshner, "A unified formulation of Gaussian vs. sparse stochastic processes—Part II: Discrete-domain theory," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 3036–3051, May 2014.
[15] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5193–5201, Nov. 2011.
[16] R. Gribonval, V. Cevher, and M. E. Davies, "Compressible distributions for high-dimensional statistics," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5016–5034, 2012.
[17] C. L. Nikias and M. Shao, *Signal Processing With Alpha-Stable Distributions and Applications*. New York, NY, USA: Wiley, 1995.
[18] E. E. Kuruoglu, W. J. Fitzgerald, and P. J. Rayner, "Near optimal detection of signals in impulsive noise modeled with a symmetric/spl alpha/-stable distribution," *IEEE Commun. Lett.*, vol. 2, no. 10, pp. 282–284, 1998.
[19] D. Middleton, "Non-gaussian noise models in signal processing for telecommunications: New methods an results for class a and class b noise models," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1129–1149, 1999.
[20] A. Achim and E. E. Kuruoglu, "Image denoising using bivariate α-stable distributions in the complex wavelet domain," *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 17–20, 2005.
[21] S. I. Resnick, "Heavy tail modeling and teletraffic data: Special invited paper," *Ann. Statist.*, vol. 25, no. 5, pp. 1805–1869, 1997.
[22] C. M. Gallagher, "A method for fitting stable autoregressive models using the autocovariation function," *Statist. Probabil. Lett.*, vol. 53, no. 4, pp. 381–390, 2001.
[23] S. Ling, "Self-weighted least absolute deviation estimation for infinite variance autoregressive models," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 67, no. 3, pp. 381–393, 2005.
[24] I. Khalidov and M. Unser, "From differential equations to the construction of new wavelet-like bases," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1256–1267, Apr. 2006.
[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, Nov. 2012, vol. 2.
[26] J. V. Stone, *Independent Component Analysis*. Cambridge, MA, USA: MIT Press, Sep. 2004.
[27] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, Nov. 1981.
[28] I. Gelfand and N. Y. Vilenkin, *Generalized Functions*. New York, NY, USA: Academic, 1964, vol. 4.
[29] M. Sahmoudi, K. Abed-Meraim, and M. Benidir, "Blind separation of impulsive alpha-stable sources using minimum dispersion criterion," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 281–284, Apr. 2005.
[30] A. R. Soltani and R. Moeanaddin, "On dispersion of stable random vectors and its application in the prediction of multivariate stable processes," *J. Appl. Probabil.*, vol. 31, no. 3, pp. 691–699, Sept. 1994.
[31] H. Stark, J. W. Woods, and H. Stark, *Probability and Random Processes With Applications to Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[32] W. D. Ray and R. M. Driver, "Further decomposition of the Karhunen-Loève series representation of stationary random process," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 6, pp. 663–668, Nov. 1970.
[33] U. S. Kamilov, P. Pad, A. Amini, and M. Unser, "MMSE estimation of sparse Lévy processes," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 137–147, Jan. 2013.

Pedram Pad (S'08) received the B.Sc. and M.Sc. degrees in electrical engineering (communications and signal processing) in 2009 and 2011, respectively, and the B.Sc. degree in mathematical sciences (pure mathematics) in 2009, all from Sharif University of Technology (SUT), Tehran, Iran. Since December 2011, he has been pursuing the Ph.D. degree with the Biomedical Imaging Group (BIG), cole Polytechnique Fdrale de Lausanne, Lausanne, Switzerland. His research interests include different aspects of information theory andsignal processing, specially applications of the former in the latter.

Michael Unser (M'89–SM'94–F'99) is professor and director of EPFL's Biomedical Imaging Group, Lausanne, Switzerland. His primary area of investigation is biomedical image processing. He is internationally recognized for his research contributions to sampling theory, wavelets, the use of splines for image processing, stochastic processes, and computational bioimaging. He has published over 250 journal papers on those topics. He is the author with P. Tafti of the book. An introduction to sparse stochastic processes, Cambridge University Press 2014.

From 1985 to 1997, he was with the Biomedical Engineering and Instrumentation Program, National Institutes of Health, Bethesda USA, conducting research on bioimaging.

Dr. Unser has held the position of associate Editor-in-Chief (2003–2005) for the IEEE Transactions on Medical Imaging. He is currently member of the editorial boards of SIAM J. Imaging Sciences, IEEE J. Selected Topics in Signal Processing, and Foundations and Trends in Signal Processing. He is the founding chair of the technical committee on Bio Imaging and Signal Processing (BISP) of the IEEE Signal Processing Society. Prof. Unser is a fellow of the IEEE (1999), an EURASIP fellow (2009), and a member of the Swiss Academy of Engineering Sciences. He is the recipient of several international prizes including three IEEE-SPS Best Paper Awards and two Technical Achievement Awards from the IEEE (2008 SPS and EMBS 2010).