Bayesian Estimation for Continuous-Time Sparse Stochastic Processes

Arash Amini, Ulugbek S. Kamilov, *Student, IEEE,* Emrah Bostan, *Student, IEEE,* Michael Unser, *Fellow, IEEE*

***Abstract*—We consider continuous-time sparse stochastic processes from which we have only a finite number of noisy/noiseless samples. Our goal is to estimate the noiseless samples (denoising) and the signal in-between (interpolation problem). By relying on tools from the theory of splines, we derive the joint a priori distribution of the samples and show how this probability density function can be factorized. The factorization enables us to tractably implement the maximum a posteriori and minimum mean-square error (MMSE) criteria as two statistical approaches for estimating the unknowns. We compare the derived statistical methods with well-known techniques for the recovery of sparse signals, such as the ℓ_1 norm and Log (ℓ_1-ℓ_0 relaxation) regularization methods. The simulation results show that, under certain conditions, the performance of the regularization techniques can be very close to that of the MMSE estimator.**

***Index Terms*—Denoising, Interpolation, Lévy Process, MAP, MMSE, Statistical Learning, Sparse Process.**

I. Introduction

The recent popularity of regularization techniques in signal and image processing is motivated by the sparse nature of real-world data. It has resulted in the development of powerful tools for many problems such as denoising, deconvolution, and interpolation. The emergence of compressed sensing, which focuses on the recovery of sparse vectors from highly under-sampled sets of measurements, is playing a key role in this context [1], [2], [3].

Assume that the signal of interest $\{s[i]\}_{i=0}^{m}$ is a finite-length discrete signal also represented by $\mathbf{s}$ as a vector) that has a sparse or almost sparse representation in some transform or analysis domain (e.g., wavelet or DCT). Assume moreover that we only have access to noisy measurements of the form $\big\{\tilde{s}[i] = s[i] + n[i]\big\}_{i=0}^{m}$, where $\big\{n[i]\big\}_{i=0}^{m}$ denotes an additive white Gaussian noise. Then, we would like to estimate $\{s[i]\}_i$. The common sparsity-promoting variational techniques rely on penalizing the sparsity in the transform/analysis domain [4], [5] by imposing

$$\big\{\hat{s}[i]\big\}_{i=0}^{m} = \arg\min_{\{s[i]\}} \big\{\|\mathbf{s} - \tilde{\mathbf{s}}\|_{\ell_2}^2 + \lambda J_{\mathrm{sparse}}(\mathbf{s})\big\}, \tag{1}$$

where $\tilde{\mathbf{s}}$ is the vector of noisy measurements, $J_{\mathrm{sparse}}(\cdot)$ is a penalty function that reflects the sparsity constraint in the transform/analysis domain and λ is a weight that is usually set based on the noise and signal powers. The choice of $J_{\mathrm{sparse}}(\cdot) = \|\cdot\|_{\ell_1}$ is one of the favorite ones in compressed sensing when $\{s[i]\}_{i=0}^{m}$ is itself sparse [6], while the use of $J_{\mathrm{sparse}}(\mathbf{s}) = TV(\mathbf{s})$, where TV stands for total variation, is a common choice for piecewise-smooth signals that have sparse derivatives [7].

Although the estimation problem for a given set of measurements is a deterministic procedure and can be handled without recourse to statistical tools, there are benefits in viewing the problem from the stochastic perspective. For instance, one can take advantage of side information about the unobserved data to establish probability laws for all or part of the data. Moreover, a stochastic framework allows one to evaluate the performance of estimation techniques and argue about their distance from the optimal estimator. The conventional stochastic interpretation of the variational method in (1) leads to the finding that $\{\hat{s}[i]\}_{i=0}^{m}$ is the maximum a posteriori (MAP) estimate of $\{s[i]\}_{i=0}^{m}$. In this interpretation, the quadratic data term is associated with the Gaussian nature of the additive noise, while the sparsifying penalty term corresponds to the a priori distribution of the sparse input. For example, the penalty $J_{\mathrm{sparse}}(\cdot) = \|\cdot\|_{\ell_1}$ is associated with the MAP estimator with Laplace prior [8], [9]. However, investigations of the compressible/sparse priors have revealed that the Laplace distribution cannot be considered as a sparse prior [10], [11], [12]. Recently in [13], it is argued that (1) is better interpreted as the minimum mean-square error (MMSE) estimator of a sparse prior.

Though the discrete stochastic models are widely adopted for sparse signals, they only approximate the continuous nature of real-world signals. The main challenge for employing continuous models is to transpose the compressibility/sparsity concepts in the continuous domain while maintaining compatibility with the discrete domain. In [14], an extended class of piecewise-smooth signals is proposed as a candidate for continuous stochastic sparse models. This class is closely related to signals with a finite rate of innovation [15]. Based on infinitely divisible distributions, a more general stochastic framework has been recently introduced in [16], [17]. There, the continuous models include Gaussian processes (such as Brownian motion), piecewise-polynomial signals, and α-stable processes as special cases. In addition, a large portion of the introduced family is considered as compressible/sparse with respect to the definition in [11] which is compatible with the discrete definition.

In this paper, we investigate the estimation problem for the samples of the continuous-time sparse models introduced in [16], [17]. We derive the a priori and a posteriori probability density functions (pdf) of the noiseless/noisy samples.

Manuscript received December 18, 2011; revised July 27, 2012.

The authors are with the Biomedical Imaging Group (BIG), École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland. Emails: {arash.amini,ulugbek.kamilov,emrah.bostan,michael.unser}@epfl.ch.

This work was supported by the European Research Center (ERC) under FUN-SP grant.

We present a practical factorization of the prior distribution which enables us to perform statistical learning for denoising or interpolation problems. In particular, we implement the optimal MMSE estimator based on the message-passing algorithm. The implementation involves discretization and convolution of pdfs, and is in general, slower than the common variational techniques. We further compare the performance of the Bayesian and variational denoising methods. Among the variational methods, we consider quadratic, TV, and Log regularization techniques. Our results show that, by matching the regularizer to the statistics, one can almost replicate the MMSE performance.

The rest of the paper is organized as follows: In Section II, we introduce our signal model which relies on the general formulation of sparse stochastic processes proposed in [16], [17]. In Section IV, we explain the techniques for obtaining the probability density functions and, we derive the estimation methods in Section III. We study the special case of Lévy processes which is of interest in many applications in Section V, and present simulation results in Section VI. Section VII concludes the paper.

II. Signal Model

In this section, we adapt the general framework of [16] to the continuous-time stochastic model studied in this paper. We follow the same notational conventions and write the input argument of the continuous-time signals/processes inside parenthesis (e.g., $s(\cdot)$) while we employ brackets (e.g., $s[\cdot]$) for discrete-time ones. Moreover, the *tilde* diacritic is used to indicate the noisy signal. Typically, $\tilde{s}[\cdot]$ represents discrete noisy samples.

In Figure 1, we give a sketch of the model. The two main parts are the continuous-time innovation process and the linear operators. The process $s(\cdot)$ is generated by applying the shaping operator L^{-1} on the innovation process w. It can be whitened back by the inverse operator L. (Since the whitening operator is of greater importance, it is represented by L while L^{-1} refers to the shaping operator.) Furthermore, the discrete observations $\tilde{s}[\cdot]$ are formed by the noisy measurements of $s(\cdot)$.

The innovation process and the linear operators have distinct implications on the resultant process s. Our model is able to handle general innovation processes that may or may not induce sparsity/compressibility. The distinction between these two cases is identified by a function $f(\omega)$ that is called the Lévy exponent, as will be discussed in Section II-A. The sparsity/compressibility of s and, consequently, of the measurements $\tilde{s}$, is inherited from the innovations and is observed in a transform domain. This domain is tied to the operator L. In this paper, we deal with operators that we represent by all-pole differential systems, tuned by acting upon the poles.

Although the model in Figure 1 is rather classical for Gaussian innovations, the investigation of non-Gaussian innovations is nontrivial. While the transition from Gaussian to non-Gaussian necessitates the reinvestigation of every definition and result, it provides us with a more general class of stochastic processes which includes compressible/sparse signals.

A. Innovation Process

Of all white processes, the Gaussian innovation is undoubtedly the one that has been investigated most thoroughly. However, it represents only a tiny fraction of the large family of white processes, which is best explored by using Gelfand's theory of generalized random processes. In his approach, unlike with the conventional point-wise definition, the stochastic process is characterized through inner products with test functions. For this purpose, one first chooses a function space $\mathcal{E}$ of test functions (e.g., the Schwartz class $\mathcal{S}$ of smooth and rapidly decaying functions). Then, one considers the random variable given by the inner product $\langle w, \varphi\rangle$, where w represents the innovation process and $\varphi \in \mathcal{E}$ [18].

***Definition** 1:* A stochastic process is called an innovation process if

1) it is stationary, *i.e.*, the random variables $\langle w, \varphi_1\rangle$ and $\langle w, \varphi_2\rangle$ are identically distributed, provided φ_2 is a shifted version of φ_1, and
2) it is white in the sense that the random variables $\langle w, \varphi_1\rangle$ and $\langle w, \varphi_2\rangle$ are independent, provided $\varphi_1, \varphi_2 \in \mathcal{E}$ are non-overlapping test functions (i.e., $\varphi_1\varphi_2 \equiv 0$).

The characteristic form of $w(\cdot)$ is defined as

$$\forall\, \varphi \in \mathcal{E}: \quad \hat{\mathscr{P}}_w(\varphi) = \mathbb{E}\{\mathrm{e}^{-\mathrm{j}\langle w, \varphi\rangle}\}, \tag{2}$$

where $\mathbb{E}\{\cdot\}$ represents the expected-value operator. The characteristic form is a powerful tool for investigating the properties of random processes. For instance, it allows one to easily infer the probability density function of the random variable $\langle w, \varphi\rangle$, or the joint densities of $\langle w, \varphi_1\rangle, \dots, \langle w, \varphi_n\rangle$. Further details regarding characteristic forms can be found in Appendix A.

The key point in Gelfand's theory is to consider the form

$$\hat{\mathscr{P}}_w(\varphi) = \exp\left(\int_{\mathbb{R}} f\big(\varphi(x)\big)\mathrm{d}x\right). \tag{3}$$

and to provide the necessary and sufficient conditions on $f(\omega)$ (the Lévy exponent) for w to define a generalized innovation process over $\mathcal{S}'$ (dual of $\mathcal{S}$). The class of admissible Lévy exponents is characterized by the Lévy-Khintchine representation theorem [19], [20] as

$$\begin{aligned} f(\omega) &= \mathrm{j}\mu\omega - \frac{\sigma^2}{2}\omega^2 \\ &+ \int_{\mathbb{R}\setminus\{0\}} \big(\mathrm{e}^{\mathrm{j}a\omega} - 1 - \mathrm{j}\omega a \mathbb{1}_{]-1,1[}(a)\big)\, v(a)\,\mathrm{d}a, \end{aligned} \tag{4}$$

where $\mathbb{1}_{\mathcal{B}}(a) = 1$ for $a \in \mathcal{B}$ and 0 otherwise, and $v(\cdot)$ (the Lévy density) is a real-valued density function that satisfies

$$\int_{\mathbb{R}\setminus\{0\}} \min(1, a^2) v(a)\,\mathrm{d}a < \infty. \tag{5}$$

In this paper, we consider only symmetric real-valued Lévy exponents (i.e., $\mu = 0$ and $v(a) = v(-a)$). Thus, the general form of (4) is reduced to

$$f(\omega) = -\frac{\sigma^2}{2}\omega^2 + \int_{\mathbb{R}\setminus\{0\}} (\cos(a\omega) - 1)\, v(a)\,\mathrm{d}a. \tag{6}$$

Next, we discuss three particular cases of (6) which are of special interest in this paper.

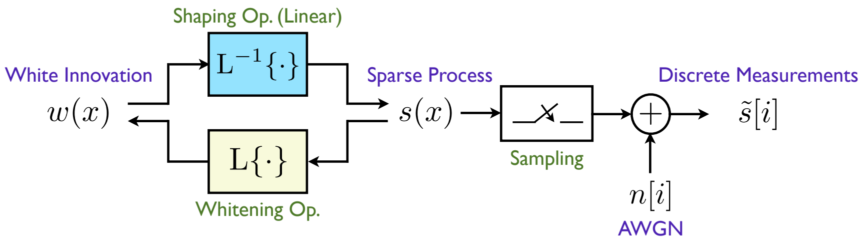


Fig. 1. Connection between the white noise $w(x)$, the sparse continuous signal $s(x)$, and the discrete measurements $\tilde{s}[i]$.

1) Gaussian Innovation: The choice $v \equiv 0$ turns (6) into

$$f_{\mathrm{G}}(\omega) = -\frac{\sigma^2}{2}\omega^2, \tag{7}$$

which implies

$$\hat{\mathscr{P}}_{w_{\mathrm{G}}}(\varphi) = \mathrm{e}^{-\frac{\sigma^2}{2}\|\varphi\|_2^2}. \tag{8}$$

This shows that the random variable $\langle w_{\mathrm{G}}, \varphi\rangle$ has a zero-mean Gaussian distribution with variance $\sigma^2\|\varphi\|_2^2$.

2) Impulsive Poisson: Let $\sigma = 0$ and $v(a) = \lambda p_a(a)$, where p_a is a symmetric probability density function. The corresponding white process is known as the impulsive Poisson innovation. By substitution in (6), we obtain

$$\begin{aligned} f_{\mathrm{IP}}(\omega) &= \lambda \int_{\mathbb{R}\setminus\{0\}} \left(\cos(a\omega) - 1\right) p_a(a)\,\mathrm{d}a \\ &= \lambda\big(\hat{p}_a(\omega) - 1\big), \end{aligned} \tag{9}$$

where $\hat{p}_a$ denotes the Fourier transform of p_a. Let $\mathbb{1}_{[0,1]}$ represents the test function that takes 1 on $[0,1]$ and 0 otherwise. Thus, if $X = \langle w_{\mathrm{IP}}, \mathbb{1}_{[0,1]}\rangle$, then for the pdf of X we know that (see Appendix A)

$$\begin{aligned} p_X(x) &= \mathcal{F}_\omega^{-1}\Big\{\mathrm{e}^{\lambda(\hat{p}_a(\omega)-1)}\Big\}(x) \\ &= \mathrm{e}^{-\lambda}\mathcal{F}_\omega^{-1}\Big\{\sum_{i=0}^{\infty}\frac{\big(\lambda\hat{p}_a(\omega)\big)^i}{i!}\Big\}(x) \\ &= \mathrm{e}^{-\lambda}\delta(x) + \sum_{i=1}^{\infty}\frac{\mathrm{e}^{-\lambda}\lambda^i}{i!}\big(\underbrace{p_a * \cdots * p_a}_{i\,\text{times}}\big)(x) \end{aligned} \tag{10}$$

It is not hard to check (see Appendix II in [14] for a proof) that this distribution matches the one that we obtain by defining

$$w(x) = \sum_{k\in\mathbb{Z}} a_k\delta(x - x_k), \tag{11}$$

where $\delta(\cdot)$ stands for the Dirac distribution, $\{x_k\}_{k\in\mathbb{Z}} \subset \mathbb{R}$ is a sequence of random Poisson points with parameter λ, and $\{a_k\}_{k\in\mathbb{Z}}$ is an independent and identically distributed (i.i.d.) sequence with probability density p_a independent of $\{x_k\}$. The sequence $\{x_k\}_{k\in\mathbb{Z}}$ is a Poisson point random sequence with parameter λ if, for all real values $a < b < c < d$, the random variables $N_1 = \big|\{x_k\}\cap[a,b]\big|$ and $N_2 = \big|\{x_k\}\cap[c,d]\big|$ are independent and N_1 (or N_2) follows the Poisson distribution with mean $\lambda(b-a)$ (or $\lambda(d-c)$), which can be written as

$$\mathrm{Prob}\{N_1 = n\} = \frac{\mathrm{e}^{-\lambda(b-a)}\big(\lambda(b-a)\big)^n}{n!}. \tag{12}$$

In [14], this type of innovation is introduced as a potential candidate for sparse processes, since all the inner products have a mass probability at $x = 0$.

3) Symmetric α-Stable: The stable laws are probability density functions that are closed under convolution. More precisely, the pdf of a random variable X is said to be stable if, for two independent and identical copies of X, namely, X_1, X_2, and for each pair of scalars $0 \le c_1, c_2$, there exists $0 \le c$ such that $c_1X_1 + c_2X_2$ has the same pdf as cX. For stable laws, it is known that $c^\alpha = c_1^\alpha + c_2^\alpha$ for some $0 < \alpha \le 2$ [21]; this is the reason why the law is indexed with α. An α-stable law which corresponds to a symmetric pdf is called symmetric α-stable. It is possible to define symmetric α-stable white processes for $0 < \alpha < 2$ by considering $\sigma = 0$ and $v(a) = \frac{c_\alpha}{|a|^{1+\alpha}}$, where $0 < c_\alpha$. From (6), we get

$$\begin{aligned} f_{\mathrm{S}}(\omega) &= c_\alpha \int_{\mathbb{R}\setminus\{0\}} \frac{\cos(a\omega) - 1}{|a|^{\alpha+1}}\,\mathrm{d}a = -2c_\alpha\int_{\mathbb{R}}\frac{\sin^2\left(\frac{a\omega}{2}\right)}{|a|^{\alpha+1}}\,\mathrm{d}a \\ &= -2c_\alpha|\omega|^\alpha\int_{\mathbb{R}}\frac{\sin^2\left(\frac{x}{2}\right)}{|x|^{\alpha+1}}\,\mathrm{d}x = -\bar{c}_\alpha|\omega|^\alpha, \end{aligned} \tag{13}$$

where $\bar{c}_\alpha$ is a positive constant. This yields

$$\hat{\mathscr{P}}_{w_{\mathrm{S}}}(\varphi) = \mathrm{e}^{-\bar{c}_\alpha\|\varphi\|_\alpha^\alpha}, \tag{14}$$

which confirms that every random variable of the form $\langle w_{\mathrm{S}}, \varphi\rangle$ has a symmetric α-stable distribution [21]. The fat-tailed distributions including α-stables for $0 < \alpha < 2$ are known to generate compressible sequences [11]. Meanwhile, the Gaussian distributions are also stable laws that correspond to the extreme value $\alpha = 2$ and have classical and well-known properties that differ fundamentally from non-Gaussian laws.

The key message of this section is that the innovation process is uniquely determined by its Lévy exponent $f(\omega)$. We shall explain in Section II-C how $f(\omega)$ affects the sparsity and compressibility properties of the process s.

B. Linear Operators

The second important component of the model is the shaping operator (the inverse of the whitening operator L) that determines the correlation structure of the process. For the generalized stochastic definition of s in Figure 1, we expect to have

$$\langle s, \varphi\rangle = \langle \mathrm{L}^{-1}w, \varphi\rangle = \langle w, \mathrm{L}^{-1*}\varphi\rangle, \tag{15}$$

where L^{-1*} represents the adjoint operator of L^{-1}. It shows that $\mathrm{L}^{-1*}\varphi$ ought to define a valid test function for the equalities in (15) to remain valid. In turn, this sets constraints on L^{-1}. The simplest choice for L^{-1} would be that of an operator which forms a continuous map from $\mathcal{S}$ into itself, but the class of such operators is not rich enough to cover the desired models in this paper. For this reason, we take advantage of a result in [16] that extends the choice of shaping operators to those L^{-1} operators for which L^{-1^*} forms a continuous mapping from $\mathcal{S}$ into L_p for some $1 \leq p$.

1) Valid Inverse Operator L^{-1}*:* In the sequel, we first explain the general requirements on the inverse of a given whitening operator L. Then, we focus on a special class of operators L and study the implications for the associated shaping operators in more details.

We assume L to be a given whitening operator, which may or may not be uniquely invertible. The minimum requirement on the shaping operator L^{-1} is that it should form a right-inverse of L (i.e., $\mathrm{L}\mathrm{L}^{-1} = \mathrm{I}$, where I is the identity operator). Furthermore, since the adjoint operator is required in (15), L^{-1} needs to be linear. This implies the existence of a kernel $h(x, \tau)$ such that

$$\mathrm{L}^{-1} w(x) = \int_{\mathbb{R}} h(x, \tau) w(\tau) \mathrm{d}\tau. \qquad (16)$$

Linear shift-invariant shaping operators are special cases that correspond to $h(x, \tau) = h(x - \tau)$. However, some of the L^{-1} operators considered in this paper are not shift-invariant.

We require the kernel h to satisfy the following three conditions:

(i) $\mathrm{L}h(x, \tau) = \delta(x - \tau)$, where L acts on the parameter x and δ is the Dirac function,
(ii) $h(x, \tau) = 0$, for $\tau > \max(0, x)$,
(iii) $(1 + |\tau|^{p-1}) \int_{\mathbb{R}} \frac{h(x,\tau)}{1+|x|^p} \mathrm{d}x$ is bounded for all $p \geq 1$.

Condition (i) is equivalent to $\mathrm{L}\mathrm{L}^{-1} = \mathrm{I}$, while (iii) is a sufficient condition studied in [16] to establish the continuity of the mapping $\mathrm{L}^{-1} : \mathcal{S} \mapsto L_p$, for all p. Condition (ii) is a constraint that we impose in this paper to simplify the statistical analysis. For $x \geq 0$, its implication is that the random variable $s(x) = \mathrm{L}^{-1} w(x)$ is fully determined by $w(\tau)$ with $\tau \leq x$, or, equivalently, it is independent of $w(\tau)$ for $\tau > x$.

From now on, we focus on differential operators L of the form $\sum_{i=0}^{n} \lambda_i \mathrm{D}^i$, where D is the first-order derivative ($\frac{\mathrm{d}}{\mathrm{d}x}$), D^0 is the identity operator (I), and λ_i are constants. With Gaussian innovations, these operators generate the autoregressive processes. An equivalent representation of L, which helps us in the analysis, is its decomposition into first-order factors as $\mathrm{L} = \lambda_n \prod_{i=1}^{n} (\mathrm{D} - r_i \mathrm{I})$. The scalars r_i are the roots of the characteristic polynomial and correspond to the poles of the inverse linear system. Here, we assume that all the poles are in the left half-plane $\Re r_i \leq 0$. This assumption helps us associate the operator L to a suitable kernel h, as shown in Appendix B.

Every differential operator L has a unique causal Green function ρ_{L} [22]. The linear shift-invariant system defined by $h(x, \tau) = \rho_{\mathrm{L}}(x - \tau)$ satisfies conditions (i)-(ii). If all the poles strictly lie in the left half-plane (i.e., $\Re r_i < 0$), due to absolute integrability of ρ_{L} (stability of the system), $h(x, \tau)$ satisfies condition (iii) as well. The definition of L^{-1} given through the kernels in Appendix B achieves both linearity and stability, while loosing shift-invariance when L contains poles on the imaginary axis. It is worth mentioning that the application of two different right-inverses of L on a given input produces results that differ only by an exponential polynomial that is in the null space of L.

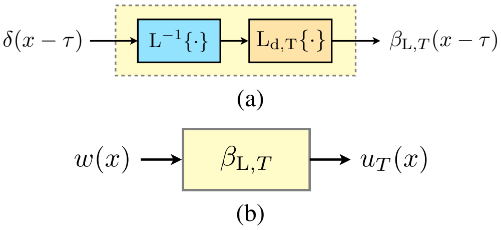

$\delta(x-\tau)$
$\mathrm{L}^{-1}\{\cdot\}$
$\mathrm{L}_{\mathrm{d,T}}\{\cdot\}$
$\beta_{\mathrm{L},T}(x-\tau)$
(a)
$w(x)$
$\beta_{\mathrm{L},T}$
$u_T(x)$
(b)

Fig. 2. (a) Linear shift-invariant operator $\mathrm{L}_{d,T}\mathrm{L}^{-1}$ and its impulse response $\beta_{\mathrm{L},T}$ (L-spline). (b) Definition of the auxiliary signal $u_T(x)$.

2) Discretization: Apart from qualifying as whitening operators, differential operators have other appealing properties such as the existence of finite-length discrete counterparts. To explain this concept, let us first consider the first-order continuous-time derivative operator D that is associated with the finite-difference filter $H(z) = 1 - z^{-1}$. This discrete counterpart is of finite length (FIR filter). Further, for any right inverse of D such as D^{-1}, the system $\mathrm{D}_{d,T}\mathrm{D}^{-1}$ is shift invariant and its impulse response is compactly supported. Here, $\mathrm{D}_{d,T}$ is the discretized operator corresponding to the sampling period T with impulse response $\big(\delta(\cdot) - \delta(\cdot - T)\big)$. It should be emphasized that the discrete counterpart $H(z)$ is a discrete-domain operator, while the discretized operator acts on continuous-domain signals. It is easy to check that this impulse response coincides with the causal B-spline of degree 0 ($\mathbb{1}_{[0,1[}$). In general, the discrete counterpart of $\mathrm{L} = \lambda_n \prod_{i=1}^{n} (\mathrm{D} - r_i \mathrm{I})$ is defined through its factors. Each $\mathrm{D} - r_i \mathrm{I}$ is associated with its discrete counterpart $H_i(z) = 1 - \mathrm{e}^{r_i} z^{-1}$ and a discretized operator given by the impulse response $\delta(\cdot) - \mathrm{e}^{r_i T} \delta(\cdot - T)$. The convolution of n such impulse responses gives rise to the impulse response of $\mathrm{L}_{d,T}$ (up to the scaling factor λ_n), which is the discretized operator of L for the sampling period T. By expanding the convolution, we obtain the form $\sum_{i=0}^{n} d_T[k] \delta(\cdot - kT)$ for the impulse response of $\mathrm{L}_{d,T}$. It is now evident that $\mathrm{L}_{d,T}$ corresponds to an FIR filter of length $(n+1)$ represented by $\{d_T[k]\}_{k=0}^{n}$ with $d_T[0] \neq 0$. Results in spline theory confirm that, for any right inverse L^{-1} of L, the operator $\mathrm{L}_{d,T}\mathrm{L}^{-1}$ is shift invariant and the support of its impulse response is contained in $[0, nT)$ [23]. The compactly supported impulse response of $\mathrm{L}_{d,T}\mathrm{L}^{-1}$, which we denote by $\beta_{\mathrm{L},T}(\cdot)$, is usually referred to as the L-spline. We define the generalized finite differences by

$$\begin{aligned} u_T(x) &= \big(\beta_{\mathrm{L},T} * w\big)(x) = (\mathrm{L}_{d,T}\mathrm{L}^{-1} w)(x) \\ &= (\mathrm{L}_{d,T}\, s)(x) = \sum_{k=0}^{n} d_T[k] s(x - kT). \qquad (17) \end{aligned}$$

We show in Figures 2 (a), (b) the definitions of $\beta_{\mathrm{L},T}(x)$ and $u_T(x)$, respectively.

C. Sparsity/Compressibility

The innovation process can be thought of as a concatenation of independent atoms. A consequence of this independence is that the process contains no redundancies. Therefore, it is incompressible under unique representation constraint.

In our framework, the role of the shaping operator L^{-1} is to generate a specific correlation structure in s by mixing the atoms of w. Conversely, the whitening operator L undoes the mixing and returns an incompressible set of data, in the sense that it maximally compresses the data. For a discretization of s corresponding to a sampling period T, the operator $\mathrm{L}_{d,T}$ mimics the role of L. It efficiently uncouples the sequence of samples and produces the generalized differences u_T, where each term depends only on a finite number of other terms. Thus, the role of $\mathrm{L}_{d,T}$ can be compared to that of converting discrete-time signals into their transform domain representation. As we explain now, the sparsity/compressibility properties of u_T are closely related to the Lévy exponent f of w.

The concept is best explained by focusing on a special class known as Lévy processes that correspond to $\beta_{\mathrm{L},T}(x) = \mathbb{1}_{[0,T[}(x)$ (see Section V for more details). By using (3) and (53), we can check that the characteristic function of the random variable u_T is given by $\mathrm{e}^{Tf(\omega)}$. When the Lévy function f is generated through a nonzero density $v(a)$ that is absolutely integrable (i.e., impulsive Poisson), the pdf of u_T associated with $\mathrm{e}^{Tf(\omega)}$ necessarily contains a mass at the origin [20] (Theorems 4.18 and 4.20). This is interpreted as sparsity when considering a finite number of measurements.

It is shown in [11], [12] that the compressibility of the measurements depends on the tail of their pdf. In simple terms, if the pdf decays at most inverse-polynomially, then, it is compressible in some corresponding L_p norm. The interesting point is that the inverse-polynomial decay of an infinite-divisible pdf is equivalent to the inverse-polynomial decay of its Lévy density $v(\cdot)$ with the same order [20] (Theorem 7.3). Therefore, an innovation process with a fat-tailed Lévy density results in processes with compressible measurements. This indicates that the slower the decay rate of $v(\cdot)$, the more compressible the measurements of s.

D. Summary of the Parameters of the Model

We now briefly review the degrees of freedom in the model and how the dependent variables are determined. The innovation process is uniquely determined by the Lévy triplet (μ, σ, v). The sparsity/compressibility of the process can be determined through the Lévy density v (see Section II-C). The values n and $\{\frac{\lambda_i}{\lambda_n}\}_{i=0}^{n-1}$, or, equivalently, the poles $\{r_i\}_{i=1}^{n}$ of the system, serve as the free parameters for the whitening operator L. As explained in Section II-B2, the taps of the FIR filter $d_T[i]$ are determined by the poles.

III. Prior Distribution

To derive statistical estimation methods such as MAP and MMSE, we need the a priori distributions. In this section, by using the generalized differences in (17), we obtain the prior distribution and factorize it efficiently. This factorization is fundamental, since it makes the implementation of the MMSE and MAP estimators tractable.

The general problem studied in this paper is to estimate $s(x)$ at $\{x = i\frac{m}{m_s}\}_{i=0}^{m_s}$ for arbitrary $m_s \in \mathbb{N}^*$ (values of the continuous process s at a uniform sampling grid with $T = \frac{m}{m_s}$), given a finite number $(m+1)$ of noisy/noiseless measurements $\{\tilde{s}[k]\}_{k=0}^{m}$ of $s(x)$ at the integers. Although we are aware that this is not equivalent to estimating the continuous process, by increasing m_s we are able to make the estimation grid as fine as desired. For piecewise-continuous signals, the limit process of refining the grid can give us access to the properties of the continuous domain.

To simplify the notations let us define

$$\left\{ \begin{array}{lcl} s_T[i] & = & s(x)|_{x=iT}, \\ u_T[i] & = & u_T(x)|_{x=iT}, \end{array} \right. \tag{18}$$

where $u_T(x)$ is defined in (17). Our goal is to derive the joint distribution of $s_T[i]$ (a priori distribution). However, the $s_T[i]$ are in general pairwise-dependent, which makes it difficult to deal with the joint distribution in high dimensions. This corresponds to a large number of samples. Meanwhile, as will be shown in Lemma 1, the sequence $\{u_T[i]\}_i$ forms a Markov chain of order $(n-1)$ that helps in factorizing the joint probability distributions, whereas $\{s_T[i]\}_i$ does not. The leading idea of this work is then that each $u_T[i]$ depends on a finite number of $u_T[j]$, $j \neq i$. It then becomes much simpler to derive the joint distribution of $\{u_T[i]\}_i$ and link it to that of $\{s_T[i]\}_i$. Lemma 1 helps us to factorize the joint pdf of $\{u_T[i]\}_i$.

Lemma 1: For $N \geq n$ and $i \geq 0$, where n is the differential order of L,

1) the random variables $\{u_T[i]\}_i$ are identically distributed,
2) the sequence $\{u_T[i]\}_i$ is a Markov chain of order $n-1$, and
3) the sample $u_T[i+N]$ is statistically independent of $u_T[i]$ and $s_T[i]$.

Proof. First note that

$$u_T[i] = \big(\beta_{\mathrm{L},T} * w\big)(x)|_{x=iT} = \langle w\,,\, \beta_{\mathrm{L},T}(iT - \cdot)\rangle. \tag{19}$$

Since $\beta_{\mathrm{L},T}(iT - \cdot)$ functions are shifts of the same function for various i and w is stationary (Definition 1), $\{u_T[i]\}_i$ are identically distributed.

Recalling that $\beta_{\mathrm{L},T}(\cdot)$ is supported on $[0, nT)$, we know that $\beta_{\mathrm{L},T}(iT - \cdot)$ and $\beta_{\mathrm{L},T}\big((i+N)T - \cdot\big)$ have no common support for $N \geq n$. Thus, due to the whiteness of w *c.f.* Definition 1), the random variables $\langle w\,,\, \beta_{\mathrm{L},T}(iT - \cdot)\rangle$ and $\langle w\,,\, \beta_{\mathrm{L},T}\big((i+N)T - \cdot\big)\rangle$ are independent. Consequently, we can write

$$\begin{aligned} & p_u\big(u_T[i+n] \mid u_T[i+n-1], u_T[i+n-2], \ldots\big) \\ & \quad = p_u\big(u_T[i+n] \mid u_T[i+n-1], \ldots, u_T[i+1]\big), \end{aligned} \tag{20}$$

which confirms that the sequence $\{u_T[i]\}_i$ forms a Markov chain of order $(n-1)$. Note that the choice of $N \geq n$ is due to the support of $\beta_{\mathrm{L},T}$. If the support of $\beta_{\mathrm{L},T}$ was infinite, it would be impossible to find j such that $u_T[i]$ and $u_T[j]$ were independent in the strict sense.

To prove the second independence property, we recall that

$$s_T[i] = s(x)\big|_{x=iT} = \int_{\mathbb{R}} h(iT,\tau)w(\tau)\mathrm{d}\tau = \langle w, h(iT,\cdot)\rangle. \quad (21)$$

Condition (ii) in Section II-B1 implies that $h(iT,\tau) = 0$ for $\tau > \max(0, iT)$. Hence, $h(iT,\tau)$ and $\beta_{\mathrm{L},T}\big((i+N)T-\cdot\big)$ have disjoint supports. Again due to whiteness of w, this implies that $u_T[i+N]$ and $s_T[i]$ are independent. ■

We now exploit the properties of $u_T[i]$ to obtain the a priori distribution of $s_T[i]$. Theorem 1, which is proved in Appendix C summarizes the main result of Section III.

***Theorem** 1:* Using the conventions of Lemma 1, for $k \geq 2n-1$ we have

$$\begin{aligned} & p_s\big(s_T[k],\ldots,s_T[0]\big) = \\ & \prod_{\theta=2n-1}^{k} \big|d_T[0]\big|\, p_u\Big(u_T[\theta] \;\Big|\; \{u_T[\theta-i]\}_{i=1}^{n-1}\Big) \\ & \times p_s\big(s_T[2n-2],\ldots,s_T[0]\big). \end{aligned} \quad (22)$$

In the definition of L^{-1} proposed in Section II-B, except when all the poles are strictly included in the left half-plane, the operator L^{-1} fails to be shift-invariant. Consequently, neither $s(x)$ nor $s_T[i]$ are stationary. An interesting consequence of Theorem 1 is that it relates the probability distribution of the non-stationary process $s_T[i]$ to that of the stationary process $u_T[i]$ plus a minimal set of transient terms.

Next, we show in Theorem 2 how the conditional probability of $u_T[i]$ can be obtained from a characteristic form. To maintain the flow of the paper, the proof is postponed to Appendix D.

***Theorem** 2:* The probability density function of $u_T[\theta]$ conditioned on $(n-1)$ previous $u_T[i]$ is given by

$$\begin{aligned} & p_u\Big(u_T[\theta] \;\Big|\; \{u_T[\theta-i]\}_{i=1}^{n-1}\Big) = \\ & \frac{\mathcal{F}^{-1}_{\{\omega_i\}}\Big\{\mathrm{e}^{I_{w,\beta_{\mathrm{L},T}}(\omega_0,\ldots,\omega_{n-1})}\Big\}\Big(\{u_T[\theta-i]\}_{i=0}^{n-1}\Big)}{\mathcal{F}^{-1}_{\{\omega_i\}}\Big\{\mathrm{e}^{I_{w,\beta_{\mathrm{L},T}}(0,\omega_1,\ldots,\omega_{n-1})}\Big\}\Big(\{u_T[\theta-i]\}_{i=1}^{n-1}\Big)}\,, \end{aligned} \quad (23)$$

where

$$\begin{aligned} & I_{w,\beta_{\mathrm{L},T}}(\omega_0,\ldots,\omega_{n-1}) \;:= \\ & \textstyle\int_{\mathbb{R}} f_w\Big(\sum_{i=0}^{n-1}\omega_i\beta_{\mathrm{L},T}(x-iT)\Big)\mathrm{d}x. \end{aligned} \quad (24)$$

IV. Signal Estimation

MMSE and MAP estimation are two common statistical paradigms for signal recovery. Since the optimal MMSE estimator is rarely realizable in practice, MAP is often used as the next best thing. In this section, in addition to applying the two methods to the proposed class of signals, we settle the question of knowing when MAP is a good approximation of MMSE.

For the estimation purpose it is convenient to assume that the sampling instances associated with $\tilde{s}[i]$ are included in the uniform sampling grid for which we want to estimate the values of s. In other words, we assume that $T = \frac{m}{m_s} = \frac{1}{n_T}$, where T is the sampling period in the definition of $s_T[\cdot]$ and n_T is a positive integer. This assumption does impose no lower bound on the resolution of the grid because we can set T arbitrarily close to zero by increasing n_T.

To simplify mathematical formulations, we use vectorial notations. We indicate the vector of noisy/noiseless measurements $\{\tilde{s}[i]\}_{i=0}^{m}$ by $\tilde{\mathbf{s}}$. The vector $\mathbf{s}_T$ stands for the hypothetical realization $\{s_T[k]\}_{k=0}^{mn_T}$ of the process on the considered grid, and $\mathbf{s}_{T,n_T}$ denotes the subset $\{s_T[in_T]\}_{i=0}^{m}$ that corresponds to the points at which we have a sample. Finally, we represent the vector of estimated values $\{\hat{s}_T[k]\}_{k=0}^{mn_T}$ by $\hat{\mathbf{s}}_T$.

A. MMSE Denoising

It is very common to evaluate the quality of an estimation method by means of the mean-square error, or SNR. In this regard, the best estimator, known as MMSE, is obtained by evaluating the posterior mean, or $\hat{\mathbf{s}}_T = \mathbb{E}\{\mathbf{s}_T \,|\, \tilde{\mathbf{s}}\}$.

For Gaussian processes, this expectation is easy to obtain, since it is equivalent to the best linear estimator [24]. However, there are only a few non-Gaussian cases where an explicit closed form of this expectation is known. In particular, if the additive noise is white and Gaussian (no restriction on the distribution of the signal) and pure denoising is desired ($T = 1$), then the MMSE estimator can be reformulated as

$$\hat{\mathbf{s}}_{\mathrm{MMSE}} = \tilde{\mathbf{s}} + \sigma_n^2 \nabla \log p_{\tilde{s}}(\mathbf{x})\big|_{\mathbf{x}=\tilde{\mathbf{s}}}\,, \quad (25)$$

where $\hat{\mathbf{s}}_{\mathrm{MMSE}}$ stands for $\hat{\mathbf{s}}_{T,\mathrm{MMSE}}$ with $T = 1$, and σ_n^2 is the variance of the noise [25], [26], [27]. Note that the pdf $p_{\tilde{s}}$, which is the result of convolving the a priori distribution p_s with the Gaussian pdf of the noise, is both continuous and differentiable.

B. MAP Estimator

Searching for the MAP estimator amounts to finding the maximizer of the distribution $p(\mathbf{s}_T \mid \tilde{\mathbf{s}})$. It is commonplace to reformulate this conditional probability in terms of the a priori distribution.

The additive discrete noise $\tilde{n}$ is white and Gaussian with the variance σ_n^2. Thus,

$$p\big(\mathbf{s}_T \,\big|\, \tilde{\mathbf{s}}\big) = p\big(\tilde{\mathbf{s}} \,\big|\, \mathbf{s}_T\big)\frac{p_s(\mathbf{s}_T)}{p_{\tilde{s}}(\tilde{\mathbf{s}})} = \frac{\mathrm{e}^{\frac{-1}{2\sigma_n^2}\|\tilde{\mathbf{s}}-\mathbf{s}_{T,n_T}\|_2^2}}{(2\pi\sigma_n^2)^{\frac{m+1}{2}}}\frac{p_s(\mathbf{s}_T)}{p_{\tilde{s}}(\tilde{\mathbf{s}})}. \quad (26)$$

In MAP estimation, we are looking for a vector $\mathbf{s}_T$ that maximizes the conditional a posteriori probability, so that (26) leads to

$$\begin{aligned} \hat{\mathbf{s}}_{T,\mathrm{MAP}} &= \arg\max_{\mathbf{s}_T}\; p\big(\mathbf{s}_T \,\big|\, \tilde{\mathbf{s}}\big) \\ &= \arg\max_{\mathbf{s}_T}\; \frac{\mathrm{e}^{\frac{-1}{2\sigma_n^2}\|\tilde{\mathbf{s}}-\mathbf{s}_{T,n_T}\|_2^2}\, p_s(\mathbf{s}_T)}{(2\pi\sigma_n^2)^{\frac{m+1}{2}}\, p_{\tilde{s}}(\tilde{\mathbf{s}})} \\ &= \arg\max_{\mathbf{s}_T}\; \mathrm{e}^{\frac{-1}{2\sigma_n^2}\|\tilde{\mathbf{s}}-\mathbf{s}_{T,n_T}\|_2^2} p_s(\mathbf{s}_T). \end{aligned} \quad (27)$$

The last equality is due to the fact that neither $(2\pi\sigma_n^2)^{\frac{m+1}{2}}$ nor $p_{\tilde{s}}(\tilde{\mathbf{s}})$ depend on the choice of $\mathbf{s}_T$. Therefore, they play no role in the maximization.

If the pdf of $\mathbf{s}_T$ is bounded, the cost function (27) can be replaced with its logarithm without changing the maximizer. The equivalent logarithmic maximization problem is given by

$$\hat{\mathbf{s}}_{T,\mathrm{MAP}} = \arg\min_{\mathbf{s}_T} \|\mathbf{s}_{T,n_T} - \tilde{\mathbf{s}}\|_2^2 - 2\sigma_n^2 \log p_s(\mathbf{s}_T). \quad (28)$$

By using the pdf factorization provided by Theorem 1, (28) is further simplified to

$$\begin{aligned}\hat{\mathbf{s}}_{T,\mathrm{MAP}} &= \arg\min_{\mathbf{s}_T} \Bigg\{ \|\mathbf{s}_{T,n_T} - \tilde{\mathbf{s}}\|_2^2 \\ &\quad -2\sigma_n^2 \sum_{k=2n-1}^{mn_T} \log p_u\big(u_T[k] \big| \{u_T[k-i]\}_{i=1}^{n-1}\big) \\ &\quad -2\sigma_n^2 \log p_s\big(s_T[2n-2], \dots, s_T[0]\big)\Bigg\}, \quad (29)\end{aligned}$$

where each $u_T[i]$ is provided by the linear combination of the elements in $\mathbf{s}_T$ found in (17). If we denote the term $\Big(-\log p_u\big(u_T[k]\big|\{u_T[k-i]\}_{i=1}^{n-1}\big)\Big)$ by $\Psi_T\big(u_T[k], \dots, u_T[k-n+1]\big)$, the MAP estimation becomes equivalent to the minimization problem

$$\begin{aligned}\hat{\mathbf{s}}_{T,\mathrm{MAP}} &= \arg\min_{\mathbf{s}_T} \Bigg\{ \|\mathbf{s}_{T,n_T} - \tilde{\mathbf{s}}\|_2^2 \\ &\quad +\lambda \sum_{k=2n-1}^{mn_T} \Psi_T\big(u_T[k], \dots, u_T[k-n+1]\big) \\ &\quad -\lambda \log p_s\big(s_T[2n-2], \dots, s_T[0]\big)\Bigg\}, \quad (30)\end{aligned}$$

where $\lambda = 2\sigma_n^2$. The interesting aspect is that the MAP estimator has the same form as (1) where the sparsity-promoting term Ψ_T in the cost function is determined by both L^{-1} and the distribution of the innovation. The well-known and successful TV regularizer corresponds to the special case where $\Psi_T(\cdot)$ is the univariate function $|\cdot|$ and the FIR filter $d_T[\cdot]$ is the finite-difference operator. In Appendix E, we show the existence of an innovation process for which the MAP estimation coincides with the TV regularization.

C. *MAP* vs *MMSE*

To have a rough comparison of MAP and MMSE, it is beneficial to reformulate the MMSE estimator in (25) as a variational problem similar to (30), thereby, expressing the MMSE solution as the minimizer of a cost function that consists of a quadratic term and a sparsity-promoting penalty term. In fact, for sparse priors, it is shown in [13] that the minimizer of a cost function involving the ℓ_1-norm penalty term approximates the MMSE estimator more accurately than the commonly considered MAP estimator. Here, we propose a different interpretation without going into technical details. From (25), it is clear that

$$\hat{\mathbf{s}}_{\mathrm{MMSE}} = \tilde{\mathbf{s}} + \sigma_n^2 \nabla \log p_{\tilde{s}}\big(\hat{\mathbf{s}}_{\mathrm{MMSE}} - \mathbf{b}_{\tilde{s}}\big), \quad (31)$$

where $\mathbf{b}_{\tilde{s}} = \sigma_n^2 \nabla \log p_{\tilde{s}}\big(\tilde{\mathbf{s}}\big)$. We can check that $\hat{\mathbf{s}}_{\mathrm{MMSE}}$ in (31) sets the gradient of the cost $J(\mathbf{s}) = \|\mathbf{s} - \tilde{\mathbf{s}}\|_2^2 - 2\sigma_n^2 \log p_{\tilde{s}}(\mathbf{s} - \mathbf{b}_{\tilde{s}})$ to zero. It suggests that

$$\hat{\mathbf{s}}_{\mathrm{MMSE}} = \arg\min_{\mathbf{s}} \|\mathbf{s} - \tilde{\mathbf{s}}\|_2^2 - 2\sigma_n^2 \log p_{\tilde{s}}(\mathbf{s} - \mathbf{b}_{\tilde{s}}). \quad (32)$$

which is similar to (28). The latter result is only valid when the cost function has a unique minimizer. Similarly to [13], it is possible to show that, under some mild conditions, this constraint is fulfilled. Nevertheless, for the qualitative comparison of MAP and MMSE, we only focus on the local properties of the cost functions that are involved. The main distinction between the cost functions in (32) and (28) is the required pdf. For MAP, we need p_s, which was shown to be factorizable by the introduction of generalized finite differences. For MMSE, we require $p_{\tilde{s}}$. Recall that $p_{\tilde{s}}$ is the result of convolving p_s with a Gaussian pdf. Thus, irrespective of the discontinuities of p_s, the function $p_{\tilde{s}}$ is smooth. However, the latter is no longer separable, which complicates the minimization task. The other difference is the offset term $\mathbf{b}_{\tilde{s}}$ in the MMSE cost function. For heavy-tail innovations such as α-stables, the convolution by the Gaussian pdf of the noise does not greatly affect p_s. In such cases, $p_{\tilde{s}}$ can be approximated by p_s fairly well, indicating that the MAP estimator suffers from a bias ($\mathbf{b}_{\tilde{s}}$). The effect of convolving p_s with a Gaussian pdf becomes more evident as p_s decays faster. In the extreme case where p_s is Gaussian, $p_{\tilde{s}}$ is also Gaussian (convolution of two Gaussians) with a different mean (which introduces another type of bias). The fact that MAP and MMSE estimators are equivalent for Gaussian innovations indicates that the two biases act in opposite directions and cancel out each other. In summary, for super-exponentially decaying innovations, MAP seems to be consistent with MMSE. For heavy-tail innovations, however, the MAP estimator is a biased version of the MMSE, where the effect of the bias is observable at high noise powers. The scenario in which MAP diverges most from MMSE might be the exponentially decaying innovations, where we have both a mismatch and a bias in the cost function, as will be confirmed in the experimental part of the paper.

V. Example: Lévy Process

To demonstrate the results and implications of the theory, we consider Lévy processes as special cases of the model in Figure 1. Lévy processes are roughly defined as processes with stationary and independent increments that start from zero. The processes are compatible with our model by setting $\mathrm{L} = \frac{\mathrm{d}}{\mathrm{d}x}$ (i.e., $n = 1$ and $r_1 = 0$), $\mathrm{L}^{-1} = \int_0^x$, or $h(x,\tau) = \mathbb{1}_{[0,\infty[}(x-\tau) - \mathbb{1}_{[0,\infty[}(-\tau)$ which imposes the boundary condition $s(0) = \int_0^0 w(\tau)d\tau = 0$. The corresponding discrete FIR filter has the two taps $d_T[0] = 1$ and $d_T[1] = -1$. The impulse response of $\mathrm{L}_{d,T}\mathrm{L}^{-1}$ is given by

$$\beta_{\mathrm{L},T}(x) = u(x) - u(x-T) = \begin{cases} 1, & 0 \le x < T \\ 0, & \text{otherwise.} \end{cases} \quad (33)$$

A. *MMSE Interpolation*

A classical problem is to interpolate the signal using noiseless samples. This corresponds to the estimation of $s_T[\theta]$ where $n_T \nmid \theta$ (n_T does not divide θ) by assuming $\tilde{s}[i] = s_T[in_T]$ $(0 \le i \le m)$. Although there is no noise in this scenario, we can still employ the MMSE criterion to estimate $s_T[\theta]$. We show that the MMSE interpolator of a Lévy process is the simple linear interpolator, irrespective of the type of innovation. To prove this, we assume $l\, n_T < \theta < (l+1)n_T$ and rewrite $s_T[\theta]$ as

$$s_T[\theta] \;=\; s_T[ln_T] + \sum_{k=ln_T+1}^{\theta} \underbrace{s_T[k] - s_T[k-1]}_{u_T[k]}. \quad (34)$$

This enables us to compute

$$\begin{aligned}\hat{s}_T[\theta] &= \mathbb{E}\big\{s_T[\theta] \;\big|\; \{s_T[in_T]\}_{i=0}^{m}\big\} \\ &= s_T[ln_T] + \mathbb{E}\Big\{ \sum_{k=ln_T+1}^{\theta} u_T[k] \Big| \{s_T[kn_T]\}_i \Big\} \quad (35)\end{aligned}$$

Since the mapping from the set $\{s_T[in_T]\}_i$ to $\{s_T[0]\} \cup \{u_1[i] = s_T[(i+1)n_T] - s_T[in_T]\}_i$ is one to one, the two sets can be used interchangeably for evaluating the conditional expectation. Thus,

$$\hat{s}_T[\theta] = s_T[ln_T] + \sum_{k=ln_T+1}^{\theta} \mathbb{E}\Big\{u_T[k] \Big| \{s_T[0]\} \cup \{u_1[i]\}_i\Big\}. \quad (36)$$

According to Lemma 1, $u_T[k]$ (for $k > 0$) is independent of $s_T[0]$ and $u_T[k^{'}]$, where $k \neq k^{'}$. By rewriting $u_1[i]$ as $\sum_{k^{'}=in_T+1}^{(i+1)n_T} u_T[k^{'}]$, we conclude that $u_1[i]$ is independent of $u_T[k]$ unless $in_T + 1 \leq k \leq (i+1)n_T$. Hence,

$$\hat{s}_T[\theta] = s_T[ln_T] + \sum_{k=ln_T+1}^{\theta} \mathbb{E}\Big\{u_T[k] \;\Big|\; u_1[l+1]\Big\}. \quad (37)$$

Since $u_1[l+1] = \sum_{i=ln_T+1}^{(l+1)n_T} u_T[i]$ and $\{u_T[i]\}_i$ is a sequence of i.i.d. random variables, the expected mean of $u_T[i]$ conditioned on $u_1[l+1]$ is the same for all i with $ln_T + 1 \leq i \leq (l+1)n_T$, which yields

$$\begin{aligned}\mathbb{E}\Big\{u_T[k]\Big|u_1[l+1]\Big\} &= \frac{1}{n_T} \sum_{i=ln_T+1}^{(l+1)n_T} \mathbb{E}\Big\{u_T[i]\Big|u_1[l+1]\Big\} \\ &= \frac{1}{n_T}\mathbb{E}\Big\{ \sum_{i=ln_T+1}^{(l+1)n_T} u_T[i]\Big|u_1[l+1]\Big\} \\ &= \frac{u_1[l+1]}{n_T}. \quad (38)\end{aligned}$$

By applying (38) to (37), we obtain

$$\begin{aligned}\hat{s}_T[\theta] &= s_T[ln_T] + \frac{\theta - ln_T}{n_T} u_1[l+1] \\ &= (1-\lambda_\theta)\, s_T[ln_T] + \lambda_\theta\, s_T[(l+1)n_T], \quad (39)\end{aligned}$$

where $\lambda_\theta = \frac{\theta - ln_T}{n_T}$. Obviously, (39) indicates a linear interpolation between the samples.

B. MAP Denoising

Since $n = 1$, the finite differences $u_T[i]$ are independent. Therefore, the conditional probabilities involved in Theorem 1 can be replaced with the simple pdf values

$$p_s\big(s_T[k], \ldots, s_T[0]\big) = p_s\big(s_T[0]\big) \prod_{\theta=1}^{k} p_u\big(u_T[\theta]\big). \quad (40)$$

In addition, since $f_w(0) = 0$, from Theorem 2, we have

$$\left\{ \begin{array}{lcl} I_{w,\beta_{\mathrm{L}},T}(\omega) &=& T f_w(\omega) \\ p_u\big(u_T[\theta]\big) &=& \mathcal{F}_\omega^{-1}\big\{\mathrm{e}^{T f_w(\omega)}\big\}\big(u_T[\theta]\big). \end{array}\right. \quad (41)$$

In the case of impulsive Poisson innovations, as shown in (10), the pdf of $u_T[i]$ has a single mass probability at $x = 0$. Hence, the MAP estimator will choose $u_T[i] = 0$ for all i, resulting in a constant signal. In other words, according to the MAP criterion and due to the boundary condition $s(0) = 0$, the optimal estimate is nothing but the trivial all-zero function. For the other types of innovations where the pdf of the increments $u_T[i]$ is bounded, or, equivalently, when the Lévy density $v(\cdot)$ is singular at the origin [20], one can reformulate the MAP estimation in the form of (30) as

$$\begin{aligned} \hat{s}_T[0] = 0 \qquad\qquad & \text{and} \\ \big\{\hat{s}_T[k]\big\}_{k=1}^{mn_T} = \; & \arg\min_{s_T[k]} \Big\{ \sum_{i=1}^{m} \big(\tilde{s}[i] - s_T[in_T]\big)^2 \\ & + \lambda \Psi_T\big(s_T[1]\big) \\ & + \lambda \sum_{k=2}^{mn_T} \Psi_T\big(s_T[k] - s_T[k-1]\big)\Big\} \quad (42)\end{aligned}$$

where $\Psi_T(\cdot) = -\log p_u(\cdot)$ and $\lambda = 2\sigma_n^2$. Because shifting the function Ψ_T with a fixed additive scalar does not change the minimizer of (42), we can modify the function to pass through the origin (i.e., $\Psi_T(0) = 0$). After having applied this modification, the function $\Psi_{T=1}$ presents itself as shown in Figure 3 for various innovation processes such as

1) Gaussian innovation: $\sigma = 1$, and $v(a) \equiv 0$, which implies

$$p_u(x) = \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}. \quad (43)$$

2) Laplace-type innovation: $\sigma = 0$, $v(a) = \frac{\mathrm{e}^{-\sqrt{\frac{2\mathrm{e}}{\pi}}|a|}}{|a|}$, which implies (see Appendix E)

$$p_u(x) = \sqrt{\frac{\mathrm{e}}{2\pi}}\mathrm{e}^{-\sqrt{\frac{2\mathrm{e}}{\pi}}|x|}. \quad (44)$$

The Lévy process of this innovation is known as the *variance gamma process* [28].

3) Cauchy innovation (α-stable with $\alpha = 1$): $\sigma = 0$, $v(a) = \frac{\sqrt{\frac{\mathrm{e}}{8\pi^3}}}{a^2}$, which implies

$$p_u(x) = \frac{\sqrt{\frac{8\mathrm{e}}{\pi}}}{\mathrm{e} + 8\pi x^2}. \quad (45)$$

The parameters of the above innovations are set such that they all lead to the same entropy value $\frac{\log(2\pi\mathrm{e})}{2} \approx 1.41$. The negative log-likelihoods of the first two innovation types resemble the ℓ_2 and ℓ_1 regularization terms. However, the curve of Ψ_T for the Cauchy innovation shows a nonconvex log-type function.

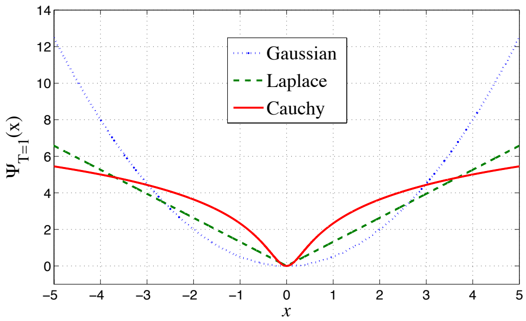


Fig. 3. The $\Psi_T(\cdot) = -\log p_u(\cdot)$ functions at $T = 1$ for Gaussian, Laplace, and Cauchy distributions. The parameters of each pdf are tuned such that they all have the same entropy and the curves are shifted to enforce them to pass through the origin.

C. MMSE Denoising

As discussed in Section IV-C, the MMSE estimator, either in the expectation form or as a minimization problem, is not separable with respect to the inputs. This is usually a critical restriction in high dimensions. Fortunately, due to the factorization of the joint a priori distribution, we can lift this restriction by employing the powerful message-passing algorithm. The method consists of representing the statistical dependencies between the parameters as a graph and finding the marginal distributions by iteratively passing the estimated pdfs along the edges [29]. The transmitted messages along the edges are also known as *beliefs*, which give rise to the alternative name of *belief propagation*. In general, the marginal distributions (beliefs) are continuous-domain objects. Hence, for computer simulations we need to discretize them.

In order to define a graphical model for a given joint probability distribution, we need to define two types of nodes: *variable* nodes that represent the input arguments for the joint pdf and *factor* nodes that portray each one of the terms in the factorization of the joint pdf. The edges in the graph are drawn only between nodes of different type and indicate the contribution of an input argument to a given factor.

For the Lévy process, we consider the joint conditional pdf $p\big(\{s_T[k]\}_k \mid \{\tilde{s}[i]\}_i\big)$ factorized as

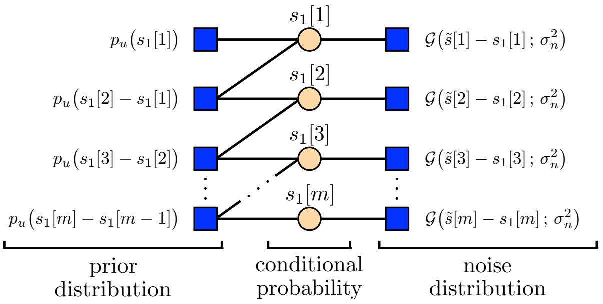
$$p\Big(\{s_T[k]\}_{k=1}^{m} \,\Big|\, \{\tilde{s}[i]\}_{i=1}^{m}\Big) = \frac{\prod_{i=1}^{m} \mathcal{G}\big(\tilde{s}[i]-s_T[i];\sigma_n^2\big) \prod_{k=1}^{mn_T} p_u(s_T[k]-s_T[k-1])}{Z}, \quad (46)$$

where $Z = p_{\tilde{s}}\left(\{\tilde{s}[i]\}_{i=1}^{m}\right)$ is a normalization constant that depends only on the noisy measurements and $\mathcal{G}$ is the Gaussian function defined as

$$\mathcal{G}\left(x;\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2\sigma^2}}. \quad (47)$$

Note that, by definition, we have $s_T[0] = 0$.

For illustration purposes, we consider the special case of pure denoising corresponding to $T = 1$. We give in Figure 4 the bipartite graph $G = (V, F, E)$ associated to the joint pdf (46). The variable nodes $V = \{1, \dots, m\}$ depicted in the middle of the graph stand for the input arguments $\{s_T[k]\}_{k=1}^{m}$. The factor nodes $F = \{1, \dots, 2m\}$ are placed at the right and left sides of the variable nodes depending on whether they represent the Gaussian factors or the $p_u(\cdot)$ factors, respectively. The set of edges $E = \{(i, a) \in V \times F\}$ also indicates a participation of the variable nodes in the corresponding factor nodes.



$p_u(s_1[1])$
$p_u(s_1[2]-s_1[1])$
$p_u(s_1[3]-s_1[2])$
$p_u(s_1[m]-s_1[m-1])$
$s_1[1]$
$s_1[2]$
$s_1[3]$
$s_1[m]$
$\mathcal{G}(\tilde{s}[1]-s_1[1]\,;\,\sigma_n^2)$
$\mathcal{G}(\tilde{s}[2]-s_1[2]\,;\,\sigma_n^2)$
$\mathcal{G}(\tilde{s}[3]-s_1[3]\,;\,\sigma_n^2)$
$\mathcal{G}(\tilde{s}[m]-s_1[m]\,;\,\sigma_n^2)$
prior distribution
conditional probability
noise distribution

Fig. 4. Factor graph for the MMSE denoising of a Lévy process. There are m variable nodes (circles) and $2m$ factor nodes (squares).

The message-passing algorithm consists of initializing the nodes of the graph with proper 1D functions and updating these functions through communications over the graph. It is desired that we eventually obtain the marginal pdf $p\big(s_1[k] \mid \{\tilde{s}[i]\}_{i=1}^{m}\big)$ on the kth variable node, which enables us to obtain the mean. The details of the messages sent over the edges and updating rules are given in [30], [31].

VI. Simulation Results

For the experiments, we consider the denoising of Lévy processes for various types of innovation, including those introduced in Section II-A and the Laplace-type innovation discussed in Appendix E. Among the heavy-tail α-stable innovations, we choose the Cauchy distribution corresponding to $\alpha = 1$. The four implemented denoising methods are

1) Linear minimum mean-square error (LMMSE) method or quadratic regularization (also known as smoothing spline [32]) defined as

$$\arg\min_{s[i]} \left\{ \|\mathbf{s} - \tilde{\mathbf{s}}\|_{\ell_2}^2 + \lambda \sum_{i=1}^{m} \big(s[i] - s[i-1]\big)^2 \right\}, \quad (48)$$

where λ should be optimized. For finding the optimum λ for given innovation statistics and a given additive-noise variance, we search for the best λ for each realization by comparing the results with the oracle estimator provided by the noiseless signal. Then, we average λ over a number of realizations to obtain a unified and realization-independent value. This procedure is repeated each time the statistics (either the innovation or the additive noise) change. For Gaussian processes, the LMMSE method coincides with both the MAP and MMSE estimators.

2) Total-variation regularization represented as

$$\arg\min_{s[i]} \left\{ \|\mathbf{s} - \tilde{\mathbf{s}}\|_{\ell_2}^2 + \lambda \sum_{i=1}^{m} \big|s[i] - s[i-1]\big| \right\}, \quad (49)$$

where λ should be optimized. The optimization process for λ is similar to the one explained for the LMMSE method.

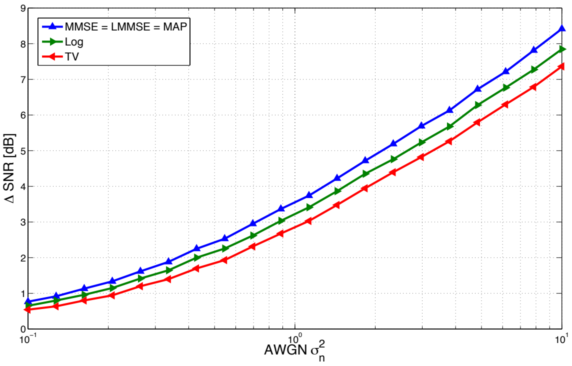


Fig. 5. SNR improvement vs. variance of the additive noise for Gaussian innovations. The denoising methods are: MMSE estimator (which is equivalent to MAP and LMMSE estimators here), Log regularization, and TV regularization.

3) Logarithmic (Log) regularization described by

$$\arg\min_{s[i]} \left\{ \|\mathbf{s}-\tilde{\mathbf{s}}\|_{\ell_2}^2 + \lambda \sum_{i=1}^{m} \log\left(1+\frac{(s[i]-s[i-1])^2}{\epsilon^2}\right) \right\}, \quad (50)$$

where λ should be optimized. The optimization process is similar to the one explained for the LMMSE method. In our experiments, we keep $\epsilon = 1$ fixed throughout the minimization steps (e.g., in the gradient-descent iterations). Unfortunately, Log is not necessarily convex, which might result in a nonconvex cost function. Hence, it is possible that gradient-descent methods get trapped in local minima rather than the desired global minimum. For heavy-tail innovations (e.g., α-stables), the Log regularizer is either the exact, or a very good approximation of, the MAP estimator.

4) Minimum mean-square error denoiser which is implemented using the message-passing technique discussed in Section V-C.

The experiments are conducted in MATLAB. We have developed a graphical user interface that facilitates the procedures of generating samples of the stochastic process and denoising them using MMSE or the variational techniques.

We show in Figure 5 the SNR improvement of a Gaussian process after denoising by the four methods. Since the LMMSE and MMSE methods are equivalent in the Gaussian case, only the MMSE curve obtained from the message-passing algorithm is plotted. As expected, the MMSE method outperforms the TV and Log regularization techniques. The counter intuitive observation is that Log, which includes a nonconvex penalty function, performs better than TV. Another advantage of the Log regularizer is that it is differentiable and quadratic around the origin.

A similar scenario is repeated in Figure 6 for the compound-Poisson innovation with $\lambda = 0.6$ and Gaussian amplitudes (zero-mean and $\sigma = 1$). As mentioned in Section V-B, since the pdf of the increments contains a mass probability at $x = 0$, the MAP estimator selects the all-zero signal as the most probable choice. In Figure 6, this trivial estimator is excluded from the comparison. It can be observed that the performance of the MMSE denoiser, which is considered to be the gold standard, is very close to that of the TV regularization method at low noise powers where the source sparsity dictates the structure. This is consistent with what was predicted in [13]. Meanwhile, it performs almost as well as the LMMSE method at large noise powers. There, the additive Gaussian noise is the dominant term and the statistics of the noisy signal is mostly determined by the Gaussian constituent, which is matched to the LMMSE method. Excluding the MMSE method, none of the other three outperforms another one for the entire range of noise.

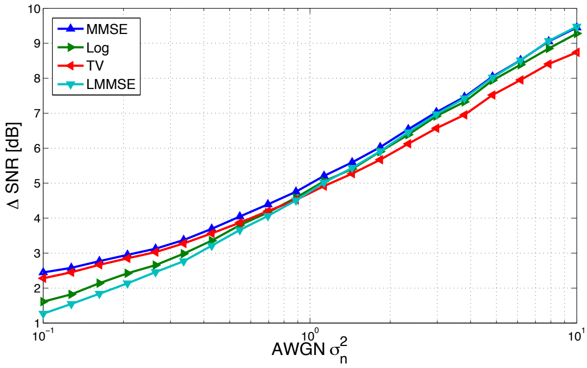


Fig. 6. SNR improvement vs. variance of the additive noise for Gaussian compound Poisson innovations. The denoising methods are: MMSE estimator, Log regularization, TV regularization, and LMMSE estimator.

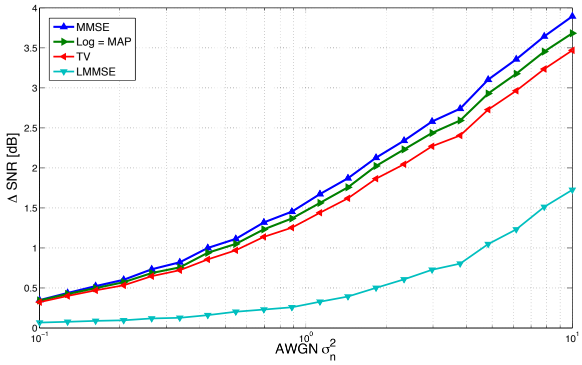


Fig. 7. SNR improvement vs. variance of the additive noise for Cauchy (α-stable with $\alpha = 1$) innovations. The denoising methods are: MMSE estimator, Log regularization (which is equivalent to MAP here), TV regularization, and LMMSE estimator.

Heavy-tail distributions such as α-stables produce sparse or compressible sequences. With high probability, their realizations consist of a few large peaks and many insignificant samples. Since the convolution of a heavy-tail pdf with a Gaussian pdf is still heavy-tail, the noisy signal looks sparse even at large noise powers. The poor performance of the LMMSE method observed in Figure 7 for Cauchy distributions confirms this characteristic. The pdf of the Cauchy distribution, given by $\frac{1}{\pi(1+x^2)}$, is in fact the symmetric α-stable distribution with

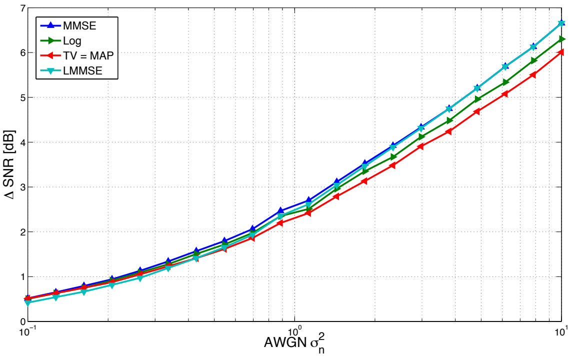


Fig. 8. SNR improvement vs. variance of the additive noise for Laplace-type innovations. The denoising methods are: MMSE estimator, Log regularization, TV regularization (which is equivalent to MAP here), and LMMSE estimator.

$\alpha = 1$. The Log regularizer corresponds to the MAP estimator of this distribution while there is no direct link between the TV regularizer and the MAP or MMSE criteria. The SNR improvement curves in Figure 7 indicate that the MMSE and Log (MAP) denoisers for this sparse process perform similarly (specially at small noise powers) and outperform the corresponding ℓ_1-norm regularizer (TV).

In the final scenario, we consider innovations with $\sigma = 0$ and $v(a) = \frac{\mathrm{e}^{-|a|}}{|a|}$. This results in finite differences obtained at $T = 1$ that follow a Laplace distribution (see Appendix E). Since the MAP denoiser for this process coincides with TV regularization, sometimes the Laplace distribution has been considered to be a sparse prior. However, it is proved in [11], [12] that the realizations of a sequence with Laplace prior are not compressible, almost surely. The curves presented in Figure 8 show that TV is a good approximation of the MMSE method only in light-noise conditions. For moderate to large noise, the LMMSE method is better than TV.

VII. Conclusion

In this paper, we studied continuous-time stochastic processes where the process is defined by applying a linear operator on a white innovation process. For specific types of innovation, the procedure results in sparse processes. We derived a factorization of the joint posterior distribution for the noisy samples of the broad family ruled by fixed-coefficient stochastic differential equations. The factorization allows us to efficiently apply statistical estimation tools. A consequence of our pdf factorization is that it gives us access to the MMSE estimator. It can then be used as a gold standard for evaluating the performance of regularization techniques. This enables us to replace the MMSE method with a more-tractable and computationally efficient regularization technique matched to the problem without compromising the performance. We then focused on Lévy processes as a special case for which we studied the denoising and interpolation problems using MAP and MMSE methods. We also compared these methods with the popular regularization techniques for the recovery of sparse signals, including the ℓ_1 norm (e.g., TV regularizer) and the Log regularization approaches. Simulation results showed that we can almost achieve the MMSE performance by tuning the regularization technique to the type of innovation and the power of the noise. We have also developed a graphical user interface in MATLAB which generates realizations of stochastic processes with various types of innovation and allows the user to apply either the MMSE or variational methods to denoise the samples[1].

Appendix A
Characteristic Forms

In Gelfand's theory of generalized random processes, the process is defined through its inner product with a space of test functions, rather than point values. For a random process w and an arbitrary test function φ chosen from a given space, the characteristic form is defined as the characteristic function of the random variable $X = \langle w, \varphi \rangle$ and given by

$$\begin{aligned} \hat{\mathscr{P}}_w(\varphi) &= \mathbb{E}\{\mathrm{e}^{-\mathrm{j}\langle w,\varphi\rangle}\} = \int_{\mathbb{R}} p_X(x)\mathrm{e}^{-\mathrm{j}x}\mathrm{d}x \\ &= \mathcal{F}_x\{p_X(x)\}(1). \end{aligned} \tag{51}$$

As an example, let w_{G} be a normalized white Gaussian noise and let φ be an arbitrary function in $L_2(\mathbb{R})$. It is well-known that $\langle w_{\mathrm{G}}, \varphi\rangle$ is a zero-mean Gaussian random variable with variance $\|\varphi\|_{L_2}^2$. Thus, in this example we have that

$$\hat{\mathscr{P}}_{w_{\mathrm{G}}}(\varphi) = \mathrm{e}^{-\frac{1}{2}\|\varphi\|_{L_2}^2}. \tag{52}$$

An interesting property of the characteristic forms is that they help determine the joint probability density functions for arbitrary finite dimensions as

$$\begin{aligned} \hat{\mathscr{P}}_w\Big(\sum_{i=1}^{k}\omega_i\varphi_i\Big) &= \mathbb{E}\Big\{\mathrm{e}^{-\mathrm{j}\langle w,\sum_{i=1}^{k}\omega_i\varphi_i\rangle}\Big\} \\ &= \mathbb{E}\Big\{\mathrm{e}^{-\mathrm{j}\sum_{i=1}^{k}\omega_i X_i}\Big\} \\ &= \mathcal{F}_{\mathbf{x}}\{p_X(\mathbf{x})\}(\omega_1,\dots,\omega_k), \end{aligned} \tag{53}$$

where $\{\varphi_i\}_i$ are test functions and $\{\omega_i\}$ are scalars. Equation (53) shows that an inverse Fourier transform of the characteristic form can yield the desired pdf. Beside joint distributions, characteristic forms are useful for generating moments too:

$$\begin{aligned} &\frac{\partial^{n_1+\dots+n_k}}{\partial\omega_1^{n_1}\cdots\partial\omega_k^{n_k}}\hat{\mathscr{P}}_w\Big(\sum_{i=1}^{k}\omega_i\varphi_i\Big)\Big|_{\omega_1=\dots=\omega_k=0} \\ &= (-\mathrm{j})^{n_1+\dots+n_k}\int_{\mathbb{R}^k} x_1^{n_1}\cdots x_k^{n_k} p_X(x_1,\dots,x_k)\mathrm{d}x_1\cdots\mathrm{d}x_k \\ &= (-\mathrm{j})^{n_1+\dots+n_k}\mathbb{E}\{x_1^{n_1}\cdots x_k^{n_k}\}. \end{aligned} \tag{54}$$

Note that the definition of random processes through characteristic forms includes the classical definition based on the point values by choosing Diracs as the test functions (if possible).

Except for the stable processes, it is usually hard to find the distributions of linear transformations of a process. However, there exists a simple relation between the characteristic forms:

[1] The GUI is available at http://bigwww.epfl.ch/amini/MATLAB_codes/SSS_GUI.zip

Let w be a random process and define $\xi = \mathrm{L}w$, where L is a linear operator. Also denote the adjoint of L by L^*. Then, one can write

$$\begin{aligned} \hat{\mathscr{P}}_\xi(\varphi) &= \mathbb{E}\{\mathrm{e}^{-\mathrm{j}\langle \mathrm{L}w,\varphi\rangle}\} = \mathbb{E}\{\mathrm{e}^{-\mathrm{j}\langle w,\mathrm{L}^*\varphi\rangle}\} \\ &= \hat{\mathscr{P}}_w(\mathrm{L}^*\varphi). \end{aligned} \tag{55}$$

Now it is easy to extract the probability distribution of ξ from its characteristic form.

Appendix B
Specification of nth-order Shaping Kernels

To show the existence of a kernel h for the nth-order differential operator $\mathrm{L} = \lambda_n \prod_{i=1}^n (\mathrm{D} - r_i\mathrm{I})$, we define

$$h_i(x,\tau) = \begin{cases} \mathrm{e}^{r_i(x-\tau)}\big(\mathbb{1}_{[0,\infty[}(x-\tau) \\ \qquad - \mathbb{1}_{[0,\infty[}(\bar{x}_i - \tau)\big), & \Re r_i = 0, \\ \mathrm{e}^{r_i(x-\tau)}\mathbb{1}_{[0,\infty[}(x-\tau), & \Re r_i < 0, \end{cases} \tag{56}$$

where $\bar{x}_i$ are nonpositive fixed real numbers. It is not hard to check that h_i satisfies the conditions (i)-(iii) for the operator $\mathrm{L}_i = \mathrm{D} - r_i\mathrm{I}$. Next, we combine h_i to form a proper kernel for L^{-1} as

$$h(x,\tau) = \frac{1}{\lambda_n}\int_{\mathbb{R}^{n-1}} \prod_{i=1}^n h_i(\tau_{i+1},\tau_i)\prod_{i=2}^n \mathrm{d}\tau_i\Big|_{\substack{\tau_1=\tau\\ \tau_{n+1}=x}}. \tag{57}$$

By relying on the fact that the h_i satisfy conditions (i)-(iii), it is possible to prove by induction that h also satisfies (i)-(iii). Here, we only provide the main idea for proving (i). We use the factorization $\mathrm{L} = \lambda_n \mathrm{L}_1 \cdots \mathrm{L}_n$ and sequentially apply every L_i on h. The starting point $i = n$ yields

$$\begin{aligned} \mathrm{L}_n h(x,\tau) &= \int_{\mathbb{R}^{n-1}} \delta(x-\tau_n)\frac{\prod_{i=1}^n h_i(\tau_{i+1},\tau_i)\mathrm{d}\tau_i}{\lambda_n h_n(\tau_{n+1},\tau_n)}\Big|_{\tau_1=\tau} \\ &= \frac{1}{\lambda_n}\int_{\mathbb{R}^{n-2}} \prod_{i=1}^{n-1} h_i(\tau_{i+1},\tau_i)\prod_{i=2}^{n-1}\mathrm{d}\tau_i\Big|_{\substack{\tau_1=\tau\\ \tau_n=x}}. \end{aligned} \tag{58}$$

Thus, $\mathrm{L}_n h(x,\tau)$ has the same form as h with n replaced by $n-1$. By continuing the same procedure, we finally arrive at $\mathrm{L}_1 h_1(x,t)$, which is equal to $\delta(x-\tau)$.

Appendix C
Proof of Theorem 1

For the sake of simplicity in the notations, for $\theta \geq n$ we define

$$\mathbf{u}[\theta] = \begin{bmatrix} u_T[\theta] \\ u_T[\theta-1] \\ \vdots \\ u_T[n] \\ s_T[n-1] \\ s_T[n-2] \\ \vdots \\ s_T[0] \end{bmatrix}. \tag{59}$$

Since the $u_T[i]$ are linear combinations of $s_T[i]$, the $\big((\theta+1)\times 1\big)$ vector $\mathbf{u}[\theta]$ can be linearly expressed in terms of $s_T[i]$ as

$$\mathbf{u}[\theta] = \mathbf{D}_{(\theta+1)\times(\theta+1)} \begin{bmatrix} s_T[\theta] \\ s_T[\theta-1] \\ \vdots \\ s_T[0] \end{bmatrix}, \tag{60}$$

where $\mathbf{D}_{(\theta+1)\times(\theta+1)}$ is an upper-triangular matrix defined by

$$\left[\begin{array}{cccccccc} d_T[0] & d_T[1] & \cdots & d_T[n] & 0 & \cdots & 0 \\ 0 & d_T[0] & \cdots & d_T[n-1] & d_T[n] & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & \cdots & 0 & d_T[0] & \cdots & d_T[n] \\ \hline \multicolumn{4}{c}{\mathbf{0}_{n\times(\theta+1-n)}} & \multicolumn{3}{c}{\mathbf{I}_{n\times n}} \end{array}\right]. \tag{61}$$

Since $d_T[0] \neq 0$, none of the diagonal elements of the upper-triangular matrix $\mathbf{D}_{(\theta+1)\times(\theta+1)}$ is zero. Thus, the matrix is invertible because $\det \mathbf{D}_{(\theta+1)\times(\theta+1)} = (d_T[0])^{\theta+1-n}$. Therefore, we have that

$$p_{s,u}\big(\mathbf{u}[\theta]\big) = \frac{p_s\big(s_T[\theta],\dots,s_T[0]\big)}{\big|d_T[0]\big|^{\theta+1-n}}. \tag{62}$$

A direct consequence of Lemma 1 is that, for $\theta \geq 2n-1$, we obtain

$$p_{s,u}\big(u_T[\theta]\,\big|\,\mathbf{u}[\theta-1]\big) = p_u\Big(u_T[\theta]\,\Big|\,\{u_T[\theta-i]\}_{i=1}^{n-1}\Big) \tag{63}$$

which, in conjunction with Bayes' rule, yields

$$\begin{aligned} \frac{p_{s,u}\big(\mathbf{u}[\theta]\big)}{p_{s,u}\big(\mathbf{u}[\theta-1]\big)} &= p_{s,u}\big(u_T[\theta]\,\big|\,\mathbf{u}[\theta-1]\big) \\ &= p_u\Big(u_T[\theta]\,\Big|\,\{u_T[\theta-i]\}_{i=1}^{n-1}\Big). \end{aligned} \tag{64}$$

By multiplying equations of the form (64) for $\theta = 2n-1,\dots,k$, we get

$$\frac{p_{s,u}\big(\mathbf{u}[k]\big)}{p_{s,u}\big(\mathbf{u}[2n-2]\big)} = \prod_{\theta=2n-1}^{k} p_u\Big(u_T[\theta]\,\Big|\,\{u_T[\theta-i]\}_{i=1}^{n-1}\Big). \tag{65}$$

It is now easy to complete the proof by substituting the numerator and denominator of the left-hand side in (65) by the equivalent forms suggested by (62).

Appendix D
Proof of Theorem 2

As developed in Appendix A, the characteristic form can be used to generate the joint probability density functions. To use (53), we need to represent $u_T[i]$ as inner-products with the white process. This is already available from (19). This yields

$$\begin{gathered} p_u\big(\{u_T[\theta-i]\}_{i=0}^k\big) = \\ \mathcal{F}^{-1}_{\{\omega_i\}}\Big\{\hat{\mathscr{P}}_w\big(\textstyle\sum_{i=0}^k \omega_i \beta_{\mathrm{L},T}(\theta T - iT - \cdot)\big)\Big\}\big(\{u_T[\theta-i]\}_{i=0}^k\big) \end{gathered} \tag{66}$$

From (3), we have

$$\hat{\mathscr{P}}_w\Big(\sum_{i=0}^{k}\omega_i\beta_{\mathrm{L},T}(\theta T - iT - \cdot)\Big)$$
$$= \mathrm{e}^{\int_{\mathbb{R}} f_w\left(\sum_{i=0}^{k}\omega_i\beta_{\mathrm{L},T}(\theta T - iT - x)\right)\mathrm{d}x}$$
$$= \mathrm{e}^{\int_{\mathbb{R}} f_w\left(\sum_{i=0}^{k}\omega_i\beta_{\mathrm{L},T}(x - iT)\right)\mathrm{d}x}. \qquad (67)$$

Using (24), it is now easy to verify that

$$\hat{\mathscr{P}}_w\Big(\sum_{i=0}^{n-1}\omega_i\beta_{\mathrm{L},T}(\theta T - iT - \cdot)\Big) = \mathrm{e}^{I_{w,\beta_{\mathrm{L},T}}(\omega_0,\dots,\omega_{n-1})},$$
$$\hat{\mathscr{P}}_w\Big(\sum_{i=1}^{n-1}\omega_i\beta_{\mathrm{L},T}(\theta T - iT - \cdot)\Big) = \mathrm{e}^{I_{w,\beta_{\mathrm{L},T}}(0,\omega_1,\dots,\omega_{n-1})}\qquad (68)$$

The only part left to mention before completing the proof is that

$$p_u\Big(u_T[\theta]\;\Big|\;\big\{u_T[\theta-i]\big\}_{i=1}^{n-1}\Big) = \frac{p_u\Big(\big\{u_T[\theta-i]\big\}_{i=0}^{n-1}\Big)}{p_u\Big(\big\{u_T[\theta-i]\big\}_{i=1}^{n-1}\Big)}. \qquad (69)$$

Appendix E
When does TV regularization meet MAP?

The TV-regularization technique is one of the successful methods in denoising. Since the TV penalty is separable with respect to first-order finite differences, its interpretation as a MAP estimator is valid only for a Lévy process. Moreover, the MAP estimator of a Lévy process coincides with TV regularization only if $\Psi_T(x) = -\log p_u(x) = \gamma|x| + \eta$, where γ and η are constants such that $\gamma > 0$. This condition implies that p_u is the Laplace pdf $p_u(x) = \frac{\gamma}{2}\mathrm{e}^{-\gamma|x|}$. This pdf is a valid distribution for the first-order finite differences of the Lévy process characterized by the innovation with $\mu = \sigma = 0$ and $v(a) = \frac{\mathrm{e}^{-\gamma|a|}}{|a|}$ because

$$\begin{aligned} f_l(\omega) &= \int_{\mathbb{R}}\big(\mathrm{e}^{\mathrm{j}\omega a} - 1\big)\frac{\mathrm{e}^{-\gamma|a|}}{|a|}\mathrm{d}a \\ &= 2\int_0^{\infty}\big(\cos(\omega a) - 1\big)\frac{\mathrm{e}^{-\gamma a}}{a}\mathrm{d}a. \end{aligned} \qquad (70)$$

Thus, we can write

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\omega}f_l(\omega) &= -2\int_0^{\infty}\sin(\omega a)\mathrm{e}^{-\gamma a}\mathrm{d}a \\ &= \frac{-2\omega}{\gamma^2+\omega^2}. \end{aligned} \qquad (71)$$

By integrating (71), we obtain that $f_l(\omega) = -\log(\gamma^2 + \omega^2) + \eta$, where η is a constant. The key point in finding this constant is the fact that $f_l(0) = 0$, which results in $f_l(\omega) = \log\frac{\gamma^2}{\gamma^2+\omega^2}$. Now, for the sampling period T, Equation (41) suggests that

$$\begin{aligned} p_u(x) &= \mathcal{F}_\omega^{-1}\{\mathrm{e}^{Tf_l(\omega)}\}(x) = \mathcal{F}_\omega^{-1}\left\{\left(\frac{\gamma^2}{\gamma^2+\omega^2}\right)^T\right\}(x) \\ &= \frac{\gamma\,|\gamma x|^{T-\frac{1}{2}}K_{T-\frac{1}{2}}(|\gamma x|)}{\sqrt{\pi}\,2^{T-\frac{1}{2}}\,\Gamma(T)}, \end{aligned} \qquad (72)$$

where $K_t(\cdot)$ is the modified Bessel function of the second kind. The latter probability density function is known as *symmetric variance-gamma* or *symm-gamma*. It is not hard to check that we obtain the desired Laplace distribution for $T = 1$. However, this value of T is the only one for which we observe this property. Should the sampling grid become finer or coarser, the MAP estimator would no longer coincide with TV regularization. We show in Figure 9 the shifted Ψ_T functions for various T values for the aforementioned innovation where $\gamma = 1$.

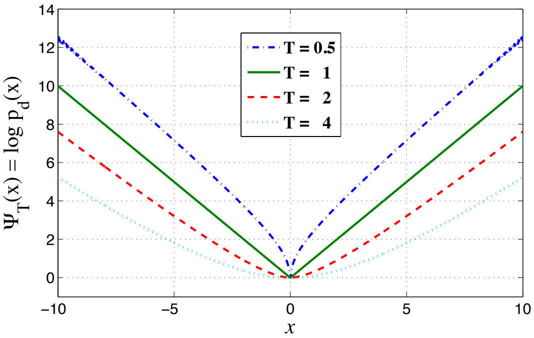


Fig. 9. The function $\Psi_T = -\log p_u$ for different values of T after enforcing the curves to pass through the origin by applying a shift. For $T = 1$, the density function p_u follows a Laplace law. Therefore, the corresponding Ψ_T is the absolute-value function.

References

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] J. L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Trans. Image Proc.*, vol. 14, no. 10, pp. 1570–1582, Oct. 2005.

[5] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Sig. Proc.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[6] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[7] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D: Nonlin. Phenom.*, vol. 60, pp. 259–268, Nov. 1992.

[8] D. P. Wipf and B. D. Rao, "Sparse Baysian learning for basis selection," *IEEE Trans. Sig. Proc.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[9] T. Park and G. Casella, "The Bayesian LASSO," *J. of the Amer. Stat. Assoc.*, vol. 103, pp. 681–686, Jun. 2008.

[10] V. Cevher, "Learning with compressible priors," in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver B.C., Canada, Dec. 8–10, 2008.

[11] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Trans. Sig. Proc.*, vol. 59, no. 11, pp. 5139–5201, Nov. 2011.

[12] R. Gribonval, V. Cevher, and M. Davies, "Compressible distributions for high-dimensional statistics," *IEEE Trans. Inform. Theo.*, vol. 58, no. 8, pp. 5016–5034, Aug. 2012.

[13] R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?" *IEEE Trans. Sig. Proc.*, vol. 59, no. 5, pp. 2405–2410, May 2011.

[14] M. Unser and P. Tafti, "Stochastic models for sparse and piecewise-smooth signals," *IEEE Trans. Sig. Proc.*, vol. 59, no. 3, pp. 989–1006, Mar. 2011.

[15] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Sig. Proc.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.

[16] M. Unser, P. Tafti, and Q. Sun, "A unified formulation of Gaussian vs. sparse stochastic processes: Part I—Continuous-Domain theory," *arXiv:1108.6150v1, http://arxiv.org/abs/1108.6150*.

[17] M. Unser, P. Tafti, A. Amini, and H. Kirshner, "A unified formulation of Gaussian vs. sparse stochastic processes: Part II—Discrete-Domain theory," *arXiv:1108.6152v1, http://arxiv.org/abs/1108.6152*.

[18] I. Gelfand and N. Y. Vilenkin, *Generalized Functions, Vol. 4. Applications of Harmonic Analysis*. New York: Academic, 1964.

[19] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*. Chapman & Hall, 1994.

[20] F. W. Steutel and K. V. Harn, *Infinite Divisibility of Probability Distributions on the Real Line*. Marcel Dekker, 2003.

[21] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*. Chapman & Hall/CRC, 1994.

[22] M. Unser and T. Blue, "Self-Similarity: Part I—Splines and operators," *IEEE Trans. Sig. Proc.*, vol. 55, no. 4, pp. 1352–1363, Apr. 2007.

[23] M. Unser and T. Blu, "Cardinal exponential splines: Part I—Theory and filtering algorithms," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1425–1438, Apr. 2005.

[24] T. Blu and M. Unser, "Self-Similarity: Part II—Optimal estimation of fractal processes," *IEEE Trans. Sig. Proc.*, vol. 55, no. 4, pp. 1364–1378, Apr. 2007.

[25] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Stat.*, vol. 9, no. 6, pp. 1135–1151, 1981.

[26] M. Raphan and E. P. Simoncelli, "Learning to be Bayesian without supervision," in *Proc. Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, Dec. 4–5, 2006, pp. 1145–1152.

[27] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. Image Proc.*, vol. 16, no. 3, pp. 593–606, Mar. 2007.

[28] D. Madan, P. Carr, and E. Chang, "The variance gamma process and option pricing," *Eur. Finance Rev.*, vol. 2, no. 1, pp. 79–105, 1998.

[29] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proc. of the IEEE*, vol. 95, no. 6, pp. 1295–1322, Jun. 2007.

[30] U. Kamilov, A. Amini, and M. Unser, "MMSE denoising of sparse Lévy processes via message passing," in *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 25-30, 2012, pp. 3637–3640.

[31] U. S. Kamilov, P. Pad, A. Amini, and M. Unser, "Mmse estimation of sparse Lévy processes," *to appear in IEEE Trans. Sig. Proc., doi: 10.1109/TSP.2012.2222394*, 2012.

[32] M. Unser and T. Blu, "Generalized smoothing splines and the optimal discretization of the Wiener filter," *IEEE Trans. Sig. Proc.*, vol. 53, no. 6, pp. 2146–2159, Jun. 2005.