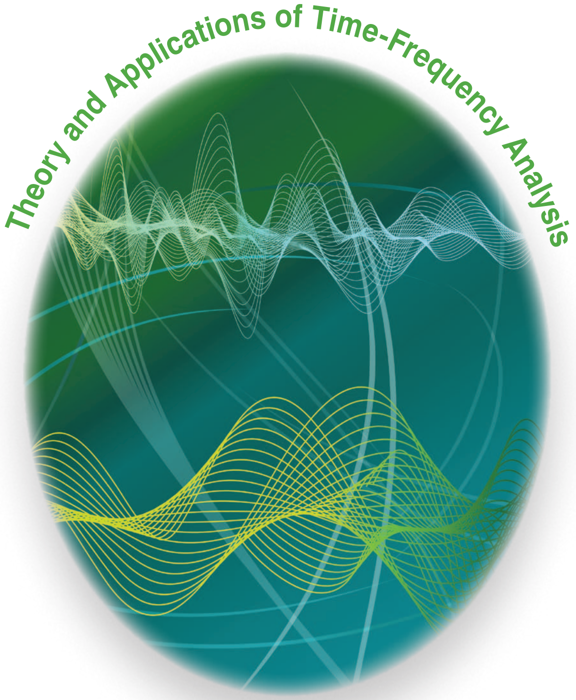
Daniele Angelosante, Georgios B. Giannakis, and Nicholas D. Sidiropoulos

Sparse Parametric Models for Robust Nonstationary Signal Analysis



Leveraging the power of sparse regression

Recent research and experimental findings, as well as technological development and commercialization efforts, suggest that even a modest amount of data can deliver superior signal modeling and reconstruction performance if sparsity is present and accounted for. Early sparsity-aware signal processing techniques have mostly targeted stationary signal analysis using offline algorithms for signal and image reconstruction from Fourier samples. On the other hand, sparsity-aware time-frequency tools for nonstationary signal analysis have recently received growing attention. In this context, sparse regression has offered a new paradigm for instantaneous frequency estimation, over classical time-frequency representations. Standard techniques for estimating model parameters from time series yield erroneous fits when, e.g., abrupt changes or outliers cause model mismatches. Accordingly, the need arises for basic research in robust processing of nonstationary parametric models that leverage sparsity to accomplish tasks such as tracking of signal variations, outlier rejection, robust

Digital Object Identifier 10.1109/MSP.2013.2267231
Date of publication: 15 October 2013

 1053-5888/13/$31.00©2013IEEE

parameter estimation, and change detection. This article aims at delineating the analytical background of sparsity-aware time-series analysis and introducing sparsity-aware robust and nonstationary parametric models to the signal processing readership, through readily appreciated applications in frequency-hopping (FH) communications and speech compression. Preliminary results strongly support the vision of seeking the "right" form of sparsity for the "right" application to enable sparsity-cognizant estimation of robust parametric models for nonstationary signal analysis.

INTRODUCTION

For over 50 years, time-series analysis has been a vibrant research topic profiting from contributions in areas as diverse as statistics, communications, econometrics, geophysics, and meteorology [3], to name a few. In the 1950s and 1960s, linear regression methods were proposed for estimation, detection, classification, estimation, and tracking based on Wiener and Kalman filtering [17], [31]. In parallel, researchers pursued approaches for (non)parametric spectral estimation. The 1970s were characterized by an increasing interest toward vector (multichannel) time series, while the 1980s brought important results in adaptive signal processing, and nonparametric interpolation techniques based on splines [30]. Major advances in the last few decades include tools for time-frequency analysis using, e.g., the Wigner–Ville distribution and the wavelet transform [10].

As evidenced by this high-level literature review, time-series analysis has slowly drifted toward nonparametric methods. One reason for this is the lack of robustness of linear parametric methods against model mismatches originating from abrupt changes of the model parameters and outliers present in the observed time series. Very recent research trends promote fully data-driven time-frequency analysis via, e.g., empirical mode decomposition (EMD) [23]. On the other hand, linear and parsimonious models have been always attractive from an implementation perspective. Such models are even better motivated for real-time systems built using simple embedded system components, where nonparametric and nonlinear methods may exhibit prohibitive complexity.

To address this performance-complexity tradeoff, a novel approach is advocated here: overcomplete linear models under which the observed signal can be sparsely represented using a suitable basis. Sparsity may arise from prior knowledge that out of a dense frequency grid, only a few frequencies can be simultaneously present. Similarly, exploiting sparsity in the model residuals can enhance robustness of classical estimators to bad data [16], [18]. Sparsity-aware time-frequency tools for nonstationary signal analysis and time-adaptive (online) algorithms have recently received growing attention [4], [13]. Indeed, when attempting to identify instantaneous frequencies, the time-frequency plane is sparsely occupied by a few trajectories. Leveraging this observation, sparse regression has offered a new paradigm for instantaneous frequency estimation over classical time-frequency representations.

To promote sparsity, the prevailing signal processing techniques employ the ℓ_1-norm of the unknown parameter vector, and, as a consequence, the estimates are nonlinear functions of the data. However, unlike classical nonlinear techniques based on, e.g., stochastic simulations, the estimates can be obtained with tractable computational complexity. This is possible because the proposed framework for robust sparsity-cognizant estimation of nonstationary parametric models is rooted in convex optimization. Historically, advanced convex optimization know-how was mostly confined to nonengineering communities, but, in the last two decades, its benefits have also permeated several engineering fields. Today, engineering practitioners can readily tap off-the-shelf convex optimization solvers (e.g., SeDuMi [1]). More recently, automatic code generation tools (e.g., CVXGEN [2]) have been developed to enable portability of convex optimization solvers to embedded systems employing simple microcontrollers and digital signal processors. This last step is crucial for bringing the power of convex optimization in the hands of practitioners.

This tutorial advocates using sparse regression as a key tool for estimating instantaneous frequencies in nonstationary time series. Its focus is on a recent sparsity-cognizant framework for time-varying (TV) frequency estimation, including both analytical modeling and computational aspects. The presented framework bridges sparse with robust regressions and employs them for the analysis of time series. Collectively, these techniques enable precise identification of parameters of complex nonstationary time series. Readily appreciated applications in FH communications and speech modeling are used to motivate and illustrate key aspects of the methodology.

ESTIMATION OF FREQUENCY-HOPPING SIGNALS

SIGNAL MODEL AND PROBLEM STATEMENT

Consider a continuous-time signal $s(t)$, which, at time $t \in [t_{k-1}, t_k]$, consists of M_k pure tones, i.e., $s(t) := \sum_{m=1}^{M_k} a_{m,k} e^{j2\pi f_{m,k} t}$, $t_{k-1} \le t < t_k$, where, $a_{m,k} \in \mathbb{C}, f_{m,k} \in [-f_{\max}, f_{\max}]$ are the complex amplitude and frequency of the mth tone in the kth system dwell $[t_{k-1}, t_k)$ (interval over which all tone frequencies and complex amplitudes remain fixed), and $t_k \in [0, T]$ is the kth system-wise hopping instant. The number of tones, M_k, can also vary with k, due to emitter (de)activation or bandwidth mismatch. The entire observation interval is $[0, T]$ (i.e., $t_0 = 0$). A noncooperative asynchronous scenario is considered, where hop timing is aperiodic and independent across transmitters. The measured continuous-time waveform $y(t)$ is corrupted by additive circularly symmetric complex white Gaussian noise $v(t)$, i.e., $y(t) := s(t) + v(t)$, $0 \le t \le T$. Let K denote the total number of system-wise hops in $[0, T]$, and $T_s := 1/(2f_{\max})$ the period with which $y(t)$ is sampled at the receiving end. The discrete-time FH signal can be written as

$$s_n := s(nT_s) = \sum_{m=1}^{M_k} a_{m,k} e^{j\omega_{m,k} n}, \quad n_{k-1} \le n < n_k, \qquad (1)$$

where $n \in \{0,1,\ldots,N-1\}$, $N-1 := \lfloor T/T_s \rfloor$, $\omega_{m,k} \in [-\pi,\pi]$, $\omega_{m,k} := 2\pi f_{m,k} T_s$, and $n_k := \lceil t_k/T_s \rceil$. Correspondingly, the discrete-time noisy observations are

$$y_n := s_n + v_n, \qquad n \in \{0,1,\ldots,N-1\}, \tag{2}$$

where $\{v_n\}$ is white and $v_n \sim \mathcal{CN}(0,\sigma^2)$. In addition, it is assumed that the sampling period is much smaller that the minimum dwell duration.

Given $\mathbf{y} := [y_0,\ldots,y_{N-1}]^T$ ($(\cdot)^T$ denotes transposition), the objective is to estimate K, $\{n_k\}$, $\{M_k\}$, $\{a_{m,k}\}$, and $\{\omega_{m,k}\}$. Since maximum-likelihood estimation of FH signal parameters is intractable, nonparametric estimators based on the spectrogram have been traditionally employed [20]. Nevertheless, the resulting estimation performance is limited by the spectrogram's intrinsic time-frequency resolution tradeoff. High-resolution time-frequency localization is possible using dynamic programming (DP) and parametric modeling [20], but complexity quickly becomes prohibitive, and performance is sensitive to model mismatch. A different and very effective approach based on sparse linear modeling is presented next.

EXPLOITING SPARSITY AND CONTINUITY TO IDENTIFY FH SIGNALS

Suppose that the true frequencies $\{\omega_{m,k}\}$ in (1) belong to a known finite set $\mathcal{W} := \{\omega_1,\ldots,\omega_P\}$ with cardinality $P \gg \max_k M_k$. Note that this is not a limiting assumption for civilian applications, e.g., Bluetooth [15], provided that Doppler is negligible. In cases where this information is not available, similar to what is used in [9] for harmonic retrieval, the set $\mathcal{W}$ can be a dense grid such that the separation between two consecutive frequencies in $\mathcal{W}$ is less than the desired resolution. Clearly, as the grid density increases with P, so does the attainable frequency resolution—what in the sparse linear regression parlance is referred to as *superresolution* [9].

With $\{\omega_{m,k}\} \subset \mathcal{W}$, the received noisy samples can be rewritten as

$$y_n = \boldsymbol{\omega}_n^T \boldsymbol{x}_n + v_n, \qquad n \in \{0,1,\ldots,N-1\}, \tag{3}$$

where $\boldsymbol{\omega}_n := [e^{j\omega_1 n},\ldots,e^{j\omega_P n}]^T$, and $\boldsymbol{x}_n := [x_{n,1},\ldots,x_{n,P}]^T \in \mathbb{C}^P$. Observe that $x_{n,p}$ represents the amplitude and phase of the pth frequency bin at time n. Since $P \gg \max_k M_k$, only a few coefficients $\{x_{n,p}\}$, representing the active frequencies at each time, are nonzero. Letting $\boldsymbol{x}^* := [\boldsymbol{x}_0^T,\ldots,\boldsymbol{x}_{N-1}^T]^T \in \mathbb{C}^{PN}$, and

$$\mathbf{w}_n := [\underbrace{\mathbf{0}_P^T,\ldots,\mathbf{0}_P^T}_{n}, \boldsymbol{\omega}_n^T, \underbrace{\mathbf{0}_P^T,\ldots,\mathbf{0}_P^T}_{N-n-1}]^T \in \mathbb{C}^{PN},$$

the model in (1), (2), and (3) can be expressed in vector-matrix form as

$$\mathbf{y} = \mathbf{W}\boldsymbol{x}^* + \mathbf{v}, \tag{4}$$

where $\mathbf{W} := [\mathbf{w}_0,\ldots,\mathbf{w}_{N-1}]^T$, and $\mathbf{v} := [v_0,\ldots,v_{N-1}]^T$. The FH signal parameters to estimate can be obtained from $\boldsymbol{x}^*$, which obeys the linear regression model in (4). Matrix $\boldsymbol{X}^* := [\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{N-1}] \in \mathbb{C}^{P\times N}$ represents the time-localized frequency content of the signal, in the same spirit as the spectrogram matrix, but with potentially much higher time-frequency resolution.

The key advantage of introducing the grid of candidate frequencies $\mathcal{W}$ is that the nonlinear parameter estimation task at hand is rendered linear [cf. (4)]. This is possible by increasing the problem dimensionality through the selection of $P \gg \max_k M_k$. Note also that as the $N \times PN$ matrix $\mathbf{W}$ is fat, the least-squares (LS) solution with minimum ℓ_2 norm, specifically, $\hat{\mathbf{x}}^{LS}_{\text{min-norm}} := \mathbf{W}^\dagger \mathbf{y}$, does not yield an accurate estimate of $\boldsymbol{x}^*$ even when the signal-to-noise ratio (SNR) is high. Improved alternatives are possible however, if one capitalizes on the fact that the unknown vector $\boldsymbol{x}^*$ exhibits the following two sparsity properties:

- *Active carrier-domain sparsity*: Only a few of the coefficients $\{x_{n,p}\}$ are nonzero, which implies that $\boldsymbol{x}^*$ in (4) is sparse
- *Differential time-domain sparsity*: Since FH is assumed slow, $x_{n+1,p} = x_{n,p}$ most of the time; hence, each row of $\boldsymbol{X}^*$ is piecewise constant. This means that adjacent row-wise differences are sparse.

Consider now the $(N-1)P \times NP$ matrix $\mathbf{D} := [\mathbf{d}_1, \mathbf{d}_1^{(1)},\ldots, \mathbf{d}_1^{((N-1)P-1)}]^T$ where

$$\mathbf{d}_1 := [-1, \underbrace{0,\ldots,0}_{P-1}, 1, \underbrace{0,\ldots,0}_{(N-1)P-1}]^T,$$

and the notation $(\cdot)^{(m)}$ represents the right cyclic shift of m positions. From the definition of $\mathbf{D}$, the $(nP+p)$th entry of $\mathbf{D}\boldsymbol{x}^*$ contains the difference $x_{n+1,p} - x_{n,p}$; hence, as mentioned earlier, $\mathbf{D}\boldsymbol{x}^*$ is a sparse vector.

Ideally, one would form a sparse and piecewise-constant estimate of $\boldsymbol{x}^*$ by solving the optimization problem

$$\check{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{C}^{NP}} \left[\frac{1}{2}\|\mathbf{y}-\mathbf{W}\mathbf{x}\|_2^2 + \mu_1\|\mathbf{x}\|_0 + \mu_2\|\mathbf{D}\mathbf{x}\|_0\right], \tag{5}$$

where $\|\mathbf{D}\mathbf{x}\|_0$ denotes the number of nonzero entries of the vector $\mathbf{D}\mathbf{x}$. The first term of the cost in (5) takes into account the observed signal, while the positive scalars μ_1 and μ_2 control the intrinsic sparsity and smoothness of the estimate, respectively. However, the problem in (5) is nonconvex and NP-hard [22].

Motivated by recent advances in variable selection [27] and compressive sampling [7], [11], the ℓ_0 (pseudo-)norm is relaxed to the convex ℓ_1-norm. Hence, the advocated formulation becomes

$$\hat{\mathbf{x}} := \arg\min_{\mathbf{x}\in\mathbb{C}^{NP}} \left[\frac{1}{2}\|\mathbf{y}-\mathbf{W}\mathbf{x}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \lambda_2\|\mathbf{D}\mathbf{x}\|_1\right]. \tag{6}$$

Large λ_1 effects sparsity, and large λ_2 effects smoothness. Since $\|\mathbf{D}\mathbf{x}\|_1 = \sum_{p=1}^{P}\sum_{n=1}^{N-1}|x_{n,p} - x_{n-1,p}|$, the second ℓ_1-norm penalty in (6) captures the sum of total variation penalties. Observe also that the Gaussian noise assumption is not necessary for (6) to be meaningful; (regularized) LS is widely used and can be motivated without recourse to Gaussianity.

The optimization problem in (6) resembles the fused Lasso [28], and it is convex because the cost comprises the sum of convex terms; hence, the cost in (6) can be minimized

ONE OR TWO FREQUENCIES—JUST BEAT IT!

In [23], an interesting phenomenon in time-frequency analysis referred to as *beating* was analyzed. Consider two closely spaced sinusoidal signals with equal amplitude, and frequencies f_1 and f_2, i.e., $s_n = \sin(2\pi f_1 n T_s) + \sin(2\pi f_2 n T_s + \phi)$. Applying prosthaphaeresis formulas, it holds that

$$\sin(2\pi f_1 n T_s) + \sin(2\pi f_2 n T_s + \phi) = 2\cos\left(2\pi \frac{f_2 - f_1}{2} n T_s + \frac{\phi}{2}\right) \sin\left(2\pi \frac{f_1 + f_2}{2} n T_s + \frac{\phi}{2}\right). \tag{S1}$$

Observe that s_n can be regarded as a single carrier at frequency $(f_1 + f_2)/2$ modulated in amplitude by the (low-frequency) sinusoidal signal $2\cos(2\pi (f_2 - f)/2\, n T_s + (\phi)/2)$. In [23], the EMD method is applied and conditions for the recovery of one or two frequencies are analyzed. Observe that the right-hand side of (S1) can be interpreted as a sparser solution than the left-hand side (one slowly TV sinusoidal component versus two stationary sinusoidal components). On the other hand, in terms of consecutive differences of the instantaneous frequencies, the left-hand side of (S1) yields sparser differences than the right-hand side.

So is it one or two frequencies? Both interpretations are valid—the "right" one depends on how one weighs sparsity versus smoothness in the instantaneous frequency (or frequencies), as reflected in the selection of λ_1 and λ_2 in (6). Reducing λ_2 while keeping λ_1 fixed will eventually tilt the balance toward a single modulated tone interpretation. This is the sparse linear regression answer to the beating phenomenon, and it is illustrated in Figure 1.

The beating phenomenon has been analyzed in several recent works dealing with signals having slowly varying spectra [23], [32]. The method presented in this article is more appropriate for FH signals. A comparison of the two approaches with respect to the beating phenomenon would be interesting, but goes beyond this article's scope.

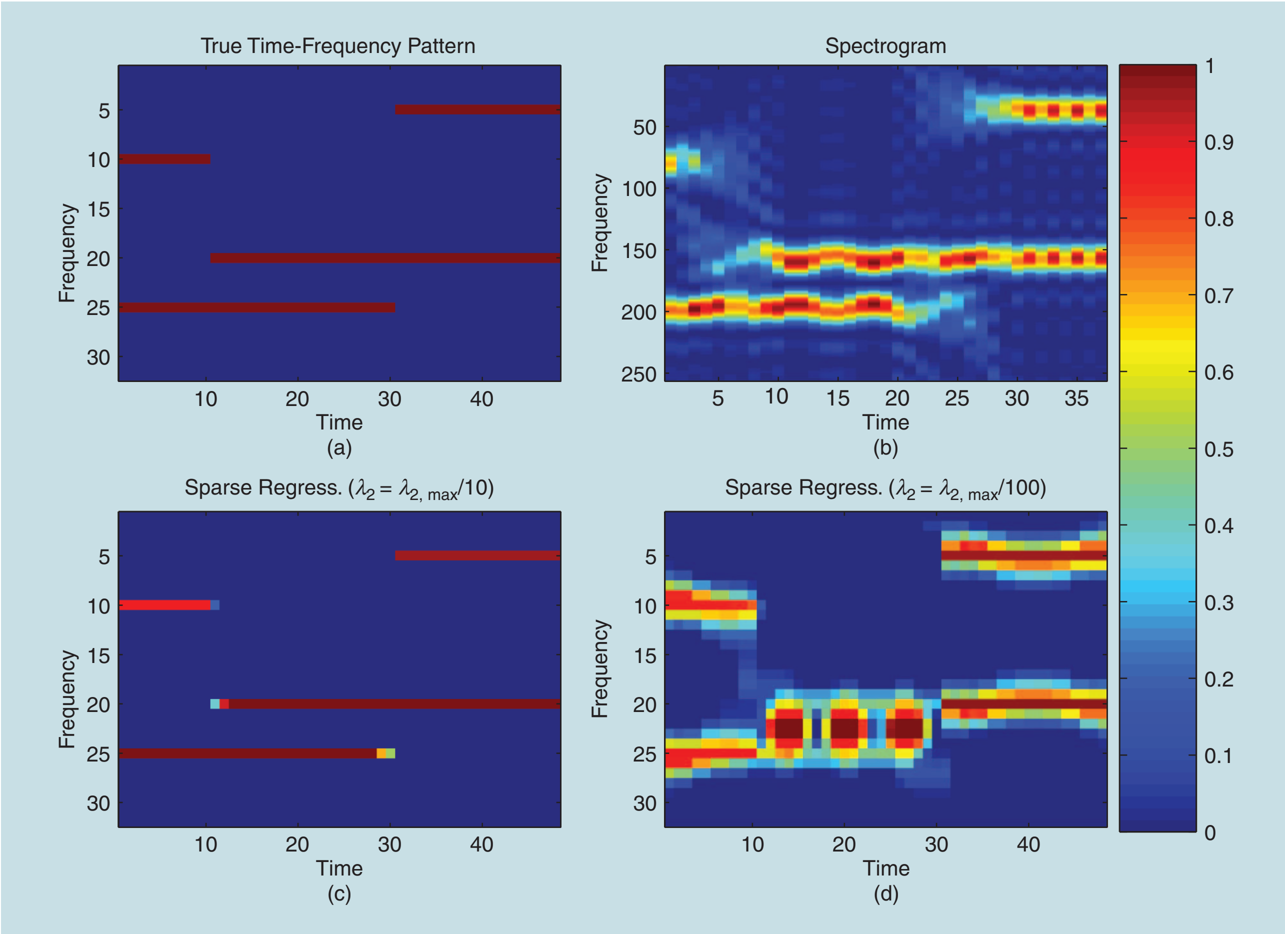
via off-the-shelf interior point solvers such as [1], which are computationally affordable for small-to-medium size problems. Large-scale problems are usually tackled via coordinate-descent solvers [29]. Nevertheless, since the nondifferentiable part in (6) is not separable coordinatewise, existing results regarding convergence of coordinate-descent solvers to a global optimum cannot be invoked [28]. An iterative algorithm to approximate the solution of the fused Lasso is developed in [28]. On the other hand, an algorithm to solve (6) exactly is derived in [6], building upon the alternating direction method of multipliers (ADMoM). Data-driven methods for tuning λ_1 and λ_1 are discussed in [6] along with generalization of (6) to polynomial phase-hopping signals (e.g., hopping chirps), and its application to wireless communications and radar. The method is quite robust to off-grid frequency mismatch, as illustrated in [6]; further improvements can be achieved by iteratively refining the frequency grid using the approach in [34].

The performance of (6) is illustrated in Figure 1 (also see "One or Two Frequencies—Just Beat It!"). The signal of interest in (1) and (3) consists of two hopping tones, while the grid of carriers is chosen to be $\mathcal{W} = \{(2p - P - 1)/(P)\pi\}_{p=1}^{P}$ with $P = 32$ and $N = 48$. The first FH tone is on the tenth carrier in the interval $[0, 9]$, and then hops to the 20th carrier during the interval $[10, 47]$. The second hopping tone occupies the 25th carrier in the interval $[0, 29]$, and the fifth carrier in the interval $[30, 47]$. The two FH signals are in-phase and have equal amplitude.

The true time-frequency pattern of the signal of interest is depicted in Figure 1(a). (Here, and in what follows, the squared modulus of the X^* entries is plotted.) The spectrogram obtained with window length $N_1 = 8$, number of frequencies $N_2 = 256$, overlap factor $L_o = 1$, and using a rectangular window is shown in Figure 1(b) at $\mathrm{SNR} := 10\log_{10}(\| x^* \|_2^2)/N\sigma^2 = 20$ dB. In Figure 1(c), the modulus of the estimate in (6) rearranged in matrix form, i.e., $\hat{X} = [\hat{x}_0, \ldots, \hat{x}_{N-1}]$, is depicted for $\lambda_1 = (\lambda_{1,\max})/10$ and $\lambda_2 = (\lambda_{2,\max})/10$, with $\lambda_{1,\max}$, $\lambda_{2,\max}$, the parameter settings [6] that recover the null solution, and the constant solution, respectively. Observe that $\hat{X}$ is a far better estimate of the true time-frequency pattern than the spectrogram. Figure 1(d), on the other hand, which depicts the solution of (6) for $\lambda_1 = (\lambda_{1,\max})/10$ and $\lambda_2 = (\lambda_{2,\max})/100$, illustrates how beating can occur at small enough λ_2. Indeed, in the dwell $[10, 29]$, a single beating frequency is identified.

JOINT SEGMENTATION AND AR MODEL IDENTIFICATION

Sum-of-exponentials models with piecewise constant parameters, such as the one advocated in the section "Estimation of Frequency-Hopping Signals," are encountered in several branches of engineering, including communications and radar. Natural signals such as speech and electroencephalogram (EEG), on the other hand, do not conform to (1). Autoregressive (AR) models have been the workhorse for parametric spectral estimation of many naturally occurring signals, since they form a dense set in the class of continuous spectra, and, in many cases, they approximate parsimoniously the spectrum of a given random process [25, Ch. 3]. While AR modeling of stationary random processes is well appreciated, a number of signals encountered in real life are nonstationary (e.g., speech signals). This justifies the growing interest toward nonstationary signal analysis and TV-AR models, which arise naturally in speech analysis due to the changing shape of the vocal tract, as well as in EEG signal analysis due to changes in the electrical activity of neurons. In the ensuing section, TV-AR models with piecewise-constant coefficients are introduced. Their identification is regarded as a sparse linear regression with grouped variables, which enables the usage of efficient algorithms.



[FIG1] The estimation of hopping complex exponentials. (a) True time-frequency pattern, (b) spectrogram, (c) sparse linear regression estimate with $\lambda_2 = \lambda_{\max}/10$, and (d) sparse linear regression estimate with $\lambda_2 = \lambda_{\max}/100$.

SIGNAL MODEL AND PROBLEM STATEMENT

Let $\{y_n\}_{n=-L}^{N}$ denote the realization of an Lth order TV-AR process obeying the discrete-time input-output relationship $y_n = \sum_{\ell=1}^{L} a_{\ell,n} y_{n-\ell} + v_n$, $n = 0, 1, \ldots, N$, where v_n denotes the zero-mean white input noise at time n with variance $\sigma^2 := \mathbb{E}[v_n^2] < \infty$, and $a_{\ell,n}$ is the ℓth TV-AR coefficient at time n. With $h_n := [y_{n-1}, y_{n-2}, \ldots, y_{n-L}]^T \in \mathbb{R}^L$ and $a_n := [a_{1,n}, a_{2,n}, \ldots, a_{L,n}]^T \in \mathbb{R}^L$, the observation model can be rewritten as

$$y_n = \mathrm{h}_n^T \mathrm{a}_n + v_n, \quad n = 0, 1, \ldots, N. \tag{7}$$

Suppose that abrupt changes in the spectrum of $\{y_n\}$ occur due to piecewise-constant changes of a_n, i.e.,

$$\mathrm{a}_n = \mathrm{a}_k, \quad n_k \le n \le n_{k+1} - 1 \tag{8}$$

for $k = 0, 1, \ldots, K$, where K denotes the number of abrupt changes in the TV-AR spectrum, and n_k the time instant of the kth abrupt change. The interval $[n_k, n_{k+1} - 1]$ is referred to as the kth *segment*. Without loss of generality, let $n_0 = 0$ and $n_{K+1} - 1 = N$.

In this context, the goal is to identify the instants $\{n_k\}_{k=1}^{K}$ where the given time series $\{y_n\}$ is split into $K+1$ (stationary) segments, and also estimate the constant AR coefficients per segment, i.e., $\{a_k\}_{k=0}^{K}$. The number of abrupt changes, specifically K, is not necessarily known.

OPTIMUM SEGMENTATION OF TV-AR PROCESSES

Regularized LS has been the workhorse approach for analyzing these kinds of nonstationary processes [19]. With μ denoting a positive tuning constant, an ℓ_0-type regularization is typically adopted to estimate jointly the change points and the AR coefficients as

$$\{\check{\mathrm{a}}_n\}_{n=0}^{N} := \arg\min_{\{\mathrm{a}_n\}_{n=0}^{N}} \left[\frac{1}{2} \sum_{n=0}^{N} (y_n - h_n^T a_n)^2 + \mu \sum_{n=1}^{N} \delta_{0_L}(a_n - a_{n-1}) \right], \tag{9}$$

where the indicator function $\delta_{0_L}(\cdot) : \mathbb{R}^L \to \{0, 1\}$ is defined as

$$\delta_{0_L}(\mathrm{a}) := \begin{cases} 0, & \text{if } \mathrm{a} = 0_L \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

The nonconvex regularization term $\sum_{n=1}^{N} \delta_{0_L}(\mathbf{a}_n - \mathbf{a}_{n-1})$ not only captures the total number of changes, but also encourages piecewise-constant $\{\check{\mathbf{a}}_n\}_{n=0}^{N}$. Clearly, the larger the μ, the smaller the total number of changes. The estimator in (9) is optimal in the maximum a posteriori (MAP) sense when the change occurrences are modeled as Bernoulli random variables, and $v_n \sim \mathcal{N}(0, \sigma^2)$ [19].

From a practical point of view, the minimization in (9) is challenging since an exhaustive search over all possible sets of change instants has to be performed. However, several techniques based on DP, simulated annealing, and interactive conditional model algorithms have been developed to evaluate (9) [19]. Even though DP approaches solve (9) in polynomial time, the computational complexity is cubic in N, which limits their applicability to signal segmentation in practice. In typical applications, N can be very large (up to several thousands), and cubic complexity cannot be afforded.

In what follows, a convex relaxation of the cost in (9) is advocated based on recent advances in sparse linear regression and compressive sampling. To this end, (9) is first reformulated into a sparse regression problem with nonconvex regularization that is subsequently relaxed through a tight convex approximation. The resulting relaxation enables remarkably accurate retrieval of change points, obtained via an efficient block-coordinate descent iteration that incurs only linear computational burden and memory storage.

EXPLOITING GROUP SPARSE COEFFICIENT CHANGES

To disclose the connections between TV-AR signal segmentation and sparse linear regression, let $\mathbf{d}_n$ denote the difference vector defined as

$$\mathbf{d}_n := \begin{cases} \mathbf{a}_n, & \text{if } n = 0 \\ \mathbf{a}_n - \mathbf{a}_{n-1}, & \text{otherwise.} \end{cases} \tag{11}$$

Using (11), the problem in (9) can be rewritten as

$$\{\check{\mathbf{d}}_n\}_{n=0}^{N} := \arg \min_{\{\mathbf{d}_n\}_{n=0}^{N}} \left[\frac{1}{2} \| \mathbf{y} - \mathbf{X}\mathbf{d} \|_2^2 + \mu \sum_{n=1}^{N} \delta_{0_L}(d_n) \right], \tag{12}$$

where $\mathbf{d} := [\mathbf{d}_0^T, \mathbf{d}_1^T, \ldots, \mathbf{d}_N^T]^T \in \mathbb{R}^{(N+1)L}$, and

$$\mathbf{X} := \begin{bmatrix} \mathbf{h}_0^T & \mathbf{0}_L^T & \cdots & \cdots & \mathbf{0}_L^T \\ \mathbf{h}_1^T & \mathbf{h}_1^T & \mathbf{0}_L^T & \cdots & \mathbf{0}_L^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{0}_L^T \\ \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T \end{bmatrix}. \tag{13}$$

Clearly, it is possible to recover $\{\mathbf{a}_n\}_{n=0}^{N}$ from $\{\mathbf{d}_n\}_{n=0}^{N}$ since $\mathbf{a}_n = \sum_{n'=0}^{n} \mathbf{d}n'$.

What makes the formulation in (12) attractive but also challenging is the nonconvex regularization term. The latter "pushes" most of the $\{\mathbf{d}_n\}_{n=1}^{N}$ vectors toward $\mathbf{0}_L$, while $\mathbf{d}_0$ is not penalized. As a consequence, the vector $\check{\mathbf{d}} := \left[\check{d}_0^T, \check{d}_1^T, \ldots, \check{d}_N^T\right]^T$ is group sparse, and the nonzero group indices correspond to the change instants of the TV-AR coefficients. Recently, a convex model selector with grouped variables was put forth by [33] and successfully applied to biostatistics and compressive sampling. It generalizes the (nongrouped) least-absolute shrinkage and selection operator (Lasso) [27] to regression problems where the unknown vector exhibits sparsity in groups; hence, its name *group Lasso*. The crux of group Lasso is to relax the regularization in (12) with a tight convex approximation. For grouped variables, it holds that the equivalent of the sparsity-promoting ℓ_1-norm is the sum of the ℓ_2-norms [33].

The group Lasso is advocated here for catching change points by estimating the difference vectors as

$$\{\hat{\mathbf{d}}_n\}_{n=0}^{N} = \arg \min_{\{\mathbf{d}_n\}_{n=0}^{N}} \left[\frac{1}{2} \| \mathbf{y} - \mathbf{X}\mathbf{d} \|_2^2 + \lambda \sum_{n=1}^{N} \| \mathbf{d}_n \|_2 \right], \tag{14}$$

where λ is a positive tuning parameter. It is known that the group Lasso regularization encourages group sparsity, i.e., $\hat{\mathbf{d}}_n = \mathbf{0}_L$ for most $n > 0$ [33]. Again, the larger the λ, the sparser the $\hat{\mathbf{d}}$.

REMARK

The role of the regularization term in (14) is illustrated next through a simple example. Select $L = 2$ for simplicity to have $\boldsymbol{d} := [d_1, d_2]^T$. Consider the family of penalties $\| \boldsymbol{d} \|_2^p = (d_1^2 + d_2^2)^{\frac{p}{2}}$, where $0 < p \leq 2$. Figure 2 depicts the penalties $\| \boldsymbol{d} \|_2^p$ for $p = 2, 1, 0.5$, and 0.1. Clearly, $\| \boldsymbol{d} \|_2^p$ is convex for $1 \leq p \leq 2$. On the other hand, $\| \boldsymbol{d} \|_2^p$ is nonconvex for $0 < p < 1$ but it comes closer to $\delta_{0_L}(\boldsymbol{d})$ as p approaches 0. Thus, it is clear that $\| \mathbf{d}_n \|_2$ is a tight convex approximation of $\delta_{0_L}(\mathbf{d}_n)$. Furthermore, $\| \mathbf{d}_n \|_2$ is nondifferentiable at $\mathbf{d}_n = \mathbf{0}_L$, which enables group Lasso to encourage group sparsity.

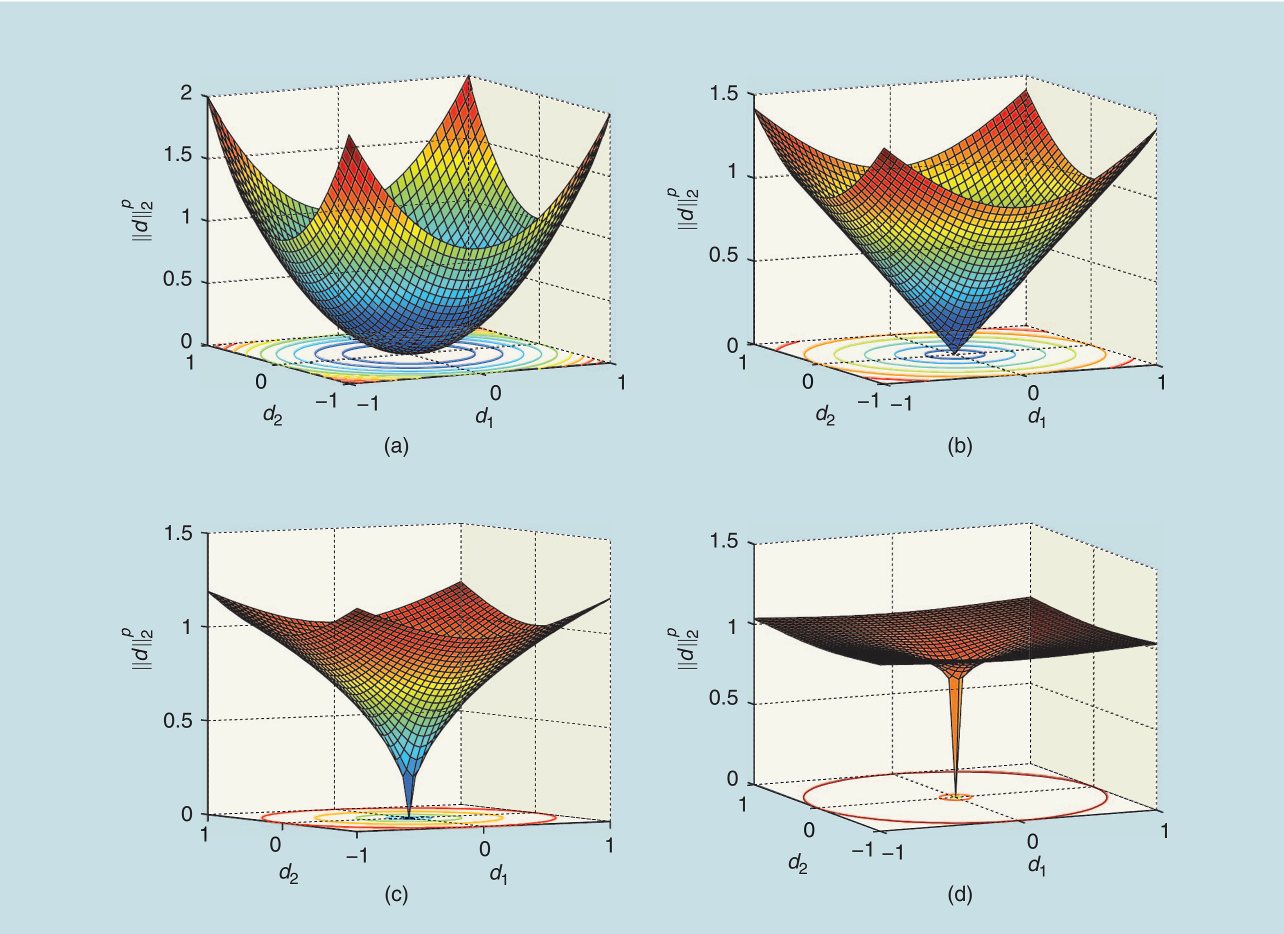
Needless to say that convexity of the regularizing functions avoids the presence of local minima, and allows for solving the resulting optimization problem efficiently. To this end, an efficient block-coordinate descent algorithm is developed in [5], with computational complexity per iteration that scales linearly with N. Furthermore, the matrix $\mathbf{X}$ does not have to be stored. Uniqueness conditions, tuning of the parameter λ, and performance enhancement are discussed in [5]. In particular, it is shown that the performance of (14) can be markedly improved if the regularization is adaptively weighted depending on interim estimates, similar to [8]. To this end, [5] advocates the smoothly clipped absolute deviation (SCAD) function, wherein the regularization term is downweighted for components that are nonzero.

The following SCAD regularization can be used with $a > 2$ in place of the group Lasso regularization:

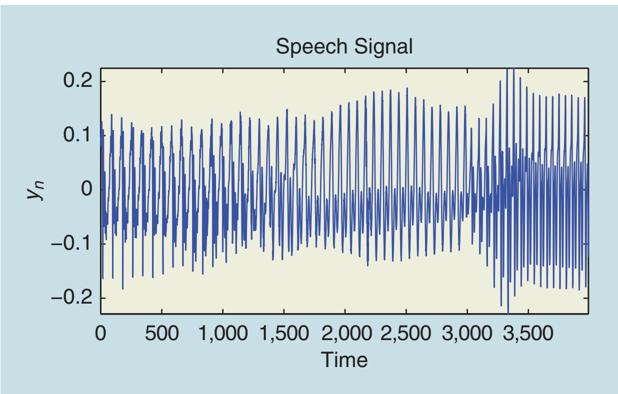
$$p_\lambda^{\text{SCAD}}(|d|) = \begin{cases} \lambda |d|, & \text{if } |d| \leq \lambda \\ -\dfrac{d^2 - 2|d|a\lambda + \lambda^2}{2(a-1)}, & \text{if } \lambda < |d| \leq a\lambda \\ \dfrac{\lambda^2}{2}(a+1), & |d| > a\lambda \end{cases} \tag{15}$$

along with the corresponding optimization problem, i.e.,

$$\{\hat{\mathbf{d}}_n\}_{n=0}^{N} = \arg \min_{\{\mathbf{d}_n\}_{n=0}^{N}} \left[\frac{1}{2} \| \mathbf{y} - \mathbf{X}\mathbf{d} \|_2^2 + \sum_{n=1}^{N} p_\lambda^{\text{SCAD}}(\| \mathbf{d}_n \|_2) \right]. \tag{16}$$



[FIG2] Regularization family $\|d\|_2^p$ for $L = 2$ and (a) $p = 2$, (b) $p = 1$, (c) $p = 0.5$, and (d) $p = 0.1$. (Figure used with permission from [5].)



[FIG3] The speech signal: /ai/-/o/. (Figure used with permission from [5].)

In the following, the performance of (14) and (16) is exemplified in the context of speech segmentation. A speech signal of 0.5 s is sampled at 8 kHz, to obtain $N + L + 1 = 4{,}000$ samples. The resulting time series depicted in Figure 3 comprises a descent diphthong /ai/ followed by an /o/. Its spectrogram, evaluated over 256-point segments with 255 samples overlap, is depicted in Figure 4. Typical speech spectra are characterized by peaks at specific frequencies called *formants*. For instance, vowel spectra are characterized by two to three formants. Therefore, to capture these spectra, typical AR model order in linear prediction coding (LPC) ranges from six to ten [21]—there is little to be gained in terms of prediction performance by using a higher model order, which does not justify the added complexity. A TV-AR model with $L = 8$ is adopted in the following. The change of vocoid in the diphthong occurs approximately at instant $n_1 = 1{,}500$, while the /o/ occurs approximately at $n_2 = 3{,}000$. Figure 5 shows the TV-AR coefficients estimated by the group Lasso and group SCAD. The group Lasso tends to declare a cloud of change points in the proximity of an actual change, while the jumps of the group SCAD estimates are very sharp. The group SCAD reveals four segments with change points at $\hat{n}_1 = 1{,}065$, $\hat{n}_2 = 1{,}606$, and $\hat{n}_3 = 2{,}993$. Clearly, the first segment corresponds to the /a/, the second, which is the shortest, to the transition of the diphthong, the third to the /i/, and the fourth to the /o/. The results in Figure 5 have been obtained in about 5 s (corresponding to 100 iterations of the coordinate descent algorithm) using MATLAB software on a Quadcore Intel Core i5 CPU running at 2.4 GHz and 6 GB of RAM.

REMARK

Recent research has shown that spiky and quasi-periodic residuals of voiced speech can be identified by using ℓ_1-norm minimization of the model residuals [14]. Our joint segmentation and TVAR system identification framework can be extended to account for sparse residuals. In fact, one can identify piecewise-constant TVAR models with sparse residuals by solving the following convex optimization problem:

$$\{\hat{\mathbf{d}}_n\}_{n=0}^{N} = \arg \min_{\{\mathbf{d}_n\}_{n=0}^{N}} \left[\frac{1}{2} \| \mathbf{y} - \mathbf{X}\mathbf{d} \|_1 + \lambda \sum_{n=1}^{N} \| \mathbf{d}_n \|_2 \right]. \qquad (17)$$

The cost in (17) resembles the one in (14), but minimization of the ℓ_1-norm of the model residuals enforces their sparsity. Performance of the method in (17) and the pursuit of efficient algorithms to solve (17) are currently under investigation.

DOUBLY ROBUST SMOOTHING

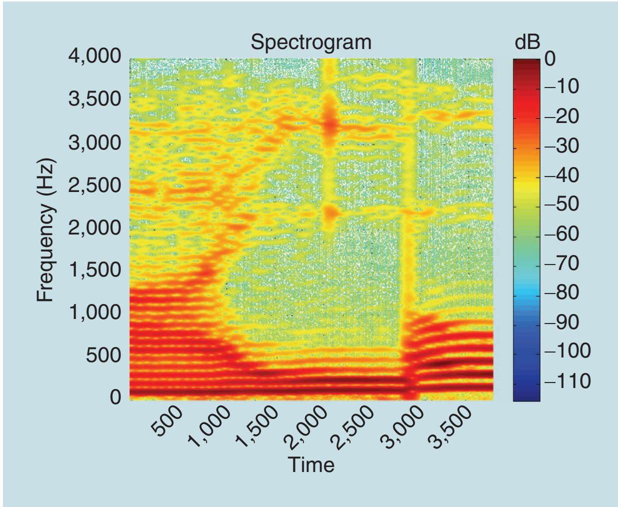
The piecewise-constant TV-AR model is important for joint speech segmentation and linear speech coding, since it allows a careful selection of the intervals to perform linear prediction. Nevertheless, most natural signals tend to exhibit spectral characteristics that are slowly time varying. Estimation of slowly TV spectra can be performed using TV-AR models that are identified via, e.g., Kalman smoothing (KS). KS has been successfully adopted to identify fine spectral characteristics of EEG and electrocardiogram (ECG) signals [26]. Nevertheless, new applications call for ever-increasing signal processing capabilities. In many cases, the signal of interest is occasionally subject to abrupt spectrum changes and bad data. It is well documented that KS falls short in this case, being sensitive to model mismatches [12]. In the ensuing section, the ordinary KS is first reviewed and recast as a convex optimization problem. KS is then made robust by introducing auxiliary (sparse) variables, which are identified along with the TV-AR model parameters.

NONROBUST KALMAN SMOOTHING

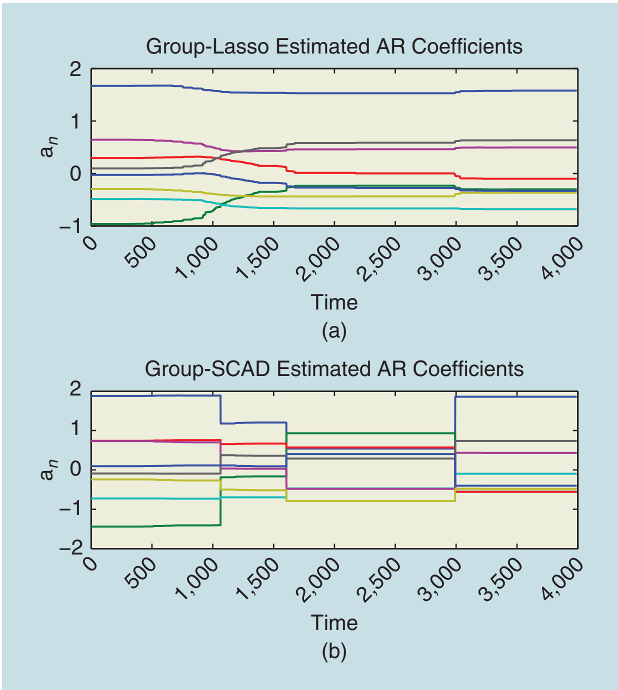
Relaxing the piecewise-constancy, it is possible to estimate slowly TV-AR processes recursively. It is assumed that the predictor coefficients change according to a random walk model

$$\mathbf{a}_n = \mathbf{a}_{n-1} + \mathbf{w}_n, \quad \text{for } n = 1, \ldots, N, \qquad (18)$$

where $\mathbf{w}_n := [w_{n,1}, \ldots, w_{n,L}]^T$, with $w_{n,\ell}$ a zero-mean, white process, with covariance $\sigma_a^2 \mathbf{I}_L$. Given observations $\{y_n\}_{n=-L}^{N}$ drawn according to (7), the optimal (in mean-squared error) linear estimation of TV-AR coefficients can be achieved via KS. KS can be regarded as the solution of the following convex optimization problem, see, e.g., [3, p. 189],



[FIG4] The spectrogram of the speech signal in Figure 3.

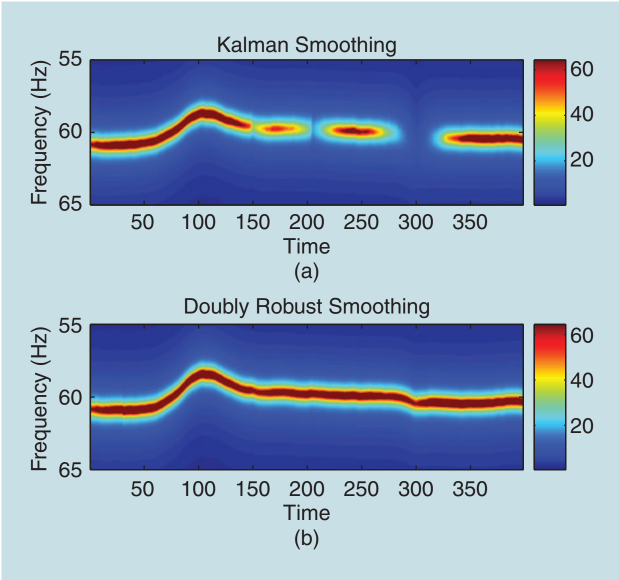


[FIG5] Estimated TV-AR coefficients by (a) group Lasso and (b) group SCAD. (Figure used with permission from [5].)

$$\{\hat{\mathbf{a}}_n^{\mathrm{KS}}\}_{n=0}^{N} = \arg \min_{\{\mathbf{a}_n\}_{n=0}^{N}} \left[\frac{1}{2\sigma_y^2} \sum_{n=0}^{N} (y_n - \mathbf{h}_n^T \mathbf{a}_n)^2 + \frac{1}{2\sigma_a^2} \sum_{n=1}^{N} \| \mathbf{a}_n - \mathbf{a}_{n-1} \|_2^2 + \frac{1}{2\sigma_a^2} \| \mathbf{a}_0 \|_2^2 \right]. \qquad (19)$$

The main advantage of KS is that the convex problem in (19) can be solved in closed-form via a first pass of the Kalman filter, followed by a backward recursion [3, p. 189].

While KS is optimal in the family of linear estimators, its performance might not be satisfactory for non-Gaussian noise,



[FIG6] The estimated TV spectrum: (a) KS and (b) doubly robust smoothing.

especially for heavy-tailed distributions of v_n, and $\mathbf{w}_n$, i.e., wherein outliers occur in the observations and the coefficients $\{\mathbf{a}_n\}$ are subject to abrupt changes [24]. In the following section, a robust smoother is presented, building upon recent advances in sparse regression.

COPING WITH OUTLIERS AND ABRUPT CHANGES

To robustify the KS in (19), unknown auxiliary variables modeling outliers and abrupt changes are introduced into the problem. The following model is adopted for doubly robust smoothing:

$$y_n = \mathbf{h}_n^T \mathbf{a}_n + v_n + o_n, \quad n = 0, 1, \ldots, N \qquad (20)$$

$$\mathbf{a}_n = \mathbf{a}_{n-1} + \mathbf{w}_n + \mathbf{c}_n, \quad n = 1, \ldots, N. \qquad (21)$$

Nonzero values in $\{o_n\}_{n=0}^N$ and $\{\mathbf{c}_n\}_{n=1}^N$ represent bad data and abrupt changes in the signal spectrum that cannot be modeled via (7) and (18). The auxiliary variables are identified along the TV-AR parameters and, since abrupt changes and bad data occur occasionally, sparsity of $\{o_n\}_{n=0}^N$ and $\{\mathbf{c}_n\}_{n=1}^N$ is imposed. Accordingly, the advocated doubly robust smoother (DRS) is [12], as can be seen in (22) in the box at the bottom of the page.

The DRS in (22) can cope with outliers jointly present in the state and in the measurements. In addition, (22) is universal because it does not require knowing the distribution of the nominal noise or the outlier vectors (in [12], data-driven criteria for the selection of ε and λ are given.)

Unlike classical KS, DRS estimates are nonlinear functions of the data. In [12], an effective solver based on the ADMoM is introduced. Closed-form expressions render the bulk of complexity per iteration comparable to that of KS, which is linear in the observation time. In practice, few iterations of the ADMoM-based algorithms are required to obtain satisfactory results. Fixed-lag DKS is also discussed in [12] for real-time applications.

To assess performance of the proposed method, a nonstationary signal with main peak around 60 Hz is acquired with sampling frequency $f_s = 200$ Hz. Two seconds, i.e., 400 samples, have been acquired. The nonstationary signal to be analyzed can be modeled as $s_n = A\sin(2\pi\,(f(n)/f_s)\,n)$. The instantaneous frequency starts at 61 Hz, then it smoothly changes to 58 Hz and stabilizes at 60 Hz. After 1.5 s, corresponding to sample index 300, the instantaneous frequency abruptly changes to 60.5 Hz. The signal $y_n := s_n + v_n + o_n$ is observed, where $v_n \sim \mathcal{N}(0, 0.05A^2)$. Bad data are represented by $\{o_n\}$, which are zero with probability 0.98, while $\{o_n\}$ are independent and identically distributed as $o_n \sim \mathcal{N}(0, A^2)$ with probability 0.02. A TV-AR model with $L = 2$ is used for the identification. Figure 6 shows the AR-estimated spectra for a realization of $\{y_n\}_{n=1}^{400}$. Bad data occur at samples indices 155 and 206. Observe that bad data strongly corrupt the KS estimates, and abrupt changes are tracked slowly. On the other hand, the DRS estimates are insensitive to bad data and abrupt changes are tracked faster.

CONCLUDING REMARKS

A set of contemporary tools was outlined in this article for time-frequency representation of nonstationary signals. These tools become available by cross-fertilizing sparsity-aware signal processing with time-series analysis. Signal models adopted in diverse branches of engineering were revisited and suitably modified to account for sparsity in different domains. It was shown that by leveraging the dual sparsity present in the frequency domain and in the time-difference domain enables accurate identification of FH signals. Piecewise-constant TV-AR models allow for joint speech segmentation and LPC. The speech segmentation problem was recast as a sparse linear regression problem that enables efficient solvers to carry out the identification task. Finally, the ordinary nonrobust KS algorithm was robustified by accounting for abrupt changes and bad data. Sparsity-aware time-frequency analysis appears to be a promising research area, if only practitioners manage to bring the appropriate form(s) of sparsity to bear on timely applications. As always, this requires a good helping of "domain art," on top of engineering principles and knowledge. Additional challenges include the pursuit of

$$[\{\hat{\mathbf{a}}_n, \hat{o}_n\}_{n=0}^N, \{\hat{\mathbf{c}}_n\}_{n=1}^N] = \arg \min_{\{\mathbf{a}_n, o_n\}_{n=0}^N, \{\mathbf{c}_n\}_{n=1}^N} \left[\frac{1}{2\sigma_y^2} \sum_{n=0}^{N} (y_n - \mathbf{h}_n^T \mathbf{a}_n - o_n)^2 + \frac{1}{2\sigma_a^2} \sum_{n=1}^{N} \| \mathbf{a}_n - \mathbf{a}_{n-1} - \mathbf{c}_n \|_2^2 + \frac{1}{2\sigma_a^2} \| \mathbf{a}_0 \|_2^2 + \varepsilon \| \mathbf{o} \|_1 + \lambda \sum_{n=1}^{N} \| \mathbf{c}_n \|_2 \right]. \qquad (22)$$

performance analysis metrics, and "frugal" algorithmic implementations that can be mass produced at a low cost.

AUTHORS

Daniele Angelosante (daniele.angelosante@ch.abb.com) received the B.Sc., Laurea Magistrale, and Ph.D. degrees in telecommunication engineering from the University of Cassino, Italy, in 2003, 2005, and 2009, respectively. From March 2009 to November 2010, he was a postdoctoral associate with the Department of Electrical and Computer Engineering of the University of Minnesota. In December 2010, he joined the corporate research center of ABB, Baden, Switzerland, as a scientist. His research interests are in the areas of embedded systems, sensors, and statistical signal processing, with emphasis on time-series analysis and compressed sensing.

Georgios B. Giannakis (georgios@umn.edu) received his Ph.D. degree from the University of Southern California in 1986. Since 1999, he has been with the University of Minnesota, where he holds the ADC chair in wireless telecommunications in the Department of Electrical and Computer Engineering and serves as director of the Digital Technology Center. His interests are in the areas of communications, networking, and statistical signal processing—subjects on which he has published more than 355 journal and 585 conference papers, 20 book chapters, two edited books, and two research monographs (h-index 105). His current research focuses on sparsity and big data analytics, cognitive networks, renewables, the power grid, and social networks. He is the (co)inventor of 21 patents and the (co)recipient of eight best paper awards from the IEEE Communications Society and IEEE Signal Processing Society (SPS). He is a Fellow of the IEEE and EURASIP and has also received technical achievement awards from SPS and EURASIP.

Nicholas D. Sidiropoulos (nikos@umn.edu) graduated from the Aristotelian University of Thessaloniki, Greece, and received his Ph.D. degree in electrical engineering from the University of Maryland in 1992. He is a professor at the University of Minnesota. His current research includes tensor decompositions, big data analytics, spectrum sensing, and cognitive radio networks. He received the SPS Best Paper Award (2001, 2007, and 2011) and the 2010 IEEE SPS Meritorious Service Award. He was an SPS Distinguished Lecturer (2008–2009) as well as the chair of the IEEE SPS Signal Processing for Communications and Networking Technical Committee. He is a Fellow of the IEEE.

REFERENCES

[1] [Online]. Available: http://sedumi.ie.lehigh.edu/

[2] [Online]. Available: http://cvxgen.com/

[3] B. Anderson and J. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice Hall, 1979.

[4] D. Angelosante, J. Bazerque, and G. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the ℓ_1-norm," *IEEE Trans. Signal Processing*, vol. 58, no. 7, pp. 3436–3447, 2010.

[5] D. Angelosante and G. Giannakis, "Group lassoing change-points in piecewise-constant AR processes," *EURASIP J. Adv. Signal Processing*, vol. 2012, no. 1, p. 70, 2012.

[6] D. Angelosante, G. Giannakis, and N. Sidiropoulos, "Estimating multiple frequency-hopping signal parameters via sparse linear regression," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5044–5056, 2010.

[7] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1-minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[9] S. Chen and D. Donoho, "Application of basis pursuit in spectrum estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1998, pp. 1865–1868.

[10] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

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

[12] S. Farahmand, G. Giannakis, and D. Angelosante, "Doubly robust smoothing of dynamical processes via outlier sparsity constraints," *IEEE Trans. Signal Processing*, vol. 59, no. 10, pp. 4529–4543, 2011.

[13] P. Flandrin and P. Borgnat, "Time-frequency energy distributions meet compressed sensing," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 2974–2982, 2010.

[14] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[15] J. Haartsen, "Bluetooth radio system," in *Encyclopedia of Telecommunications*. Hoboken, NJ: Wiley, 2003.

[16] P. Huber and E. Ronchetti, *Robust Statistics*. Hoboken, NJ: Wiley, 2009.

[17] R. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[18] V. Kekatos, and G. B. Giannakis, "From sparse signals to sparse residuals for robust sensing," *IEEE Trans. Signal Processing*, vol. 59, no. 7, pp. 3355–3368, July 2011.

[19] M. Lavielle, "Optimal segmentation of random processes," *IEEE Trans. Signal Processing*, vol. 46, no. 5, pp. 1365–1373, 1998.

[20] X. Liu, N. Sidiropoulos, and A. Swami, "Blind high-resolution localization and tracking of multiple frequency hopped signals," *IEEE Trans. Signal Processing*, vol. 50, no. 4, pp. 889–901, 2002.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[22] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[23] G. Rilling and P. Flandrin, "One or two frequencies? The empirical mode decomposition answers," *IEEE Trans. Signal Processing*, vol. 56, no. 1, pp. 85–95, 2008.

[24] I. Schick and S. Mitter, "Robust recursive estimation in the presence of heavy-tailed observation noise," *Ann. Stat.*, vol. 22, no. 2, pp. 1045–1080, 1994.

[25] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1997.

[26] M. Tarvainen, J. Hiltunen, P. Ranta-Aho, and P. Karjalainen, "Estimation of nonstationary EEG with Kalman smoother approach: An application to event-related synchronization (ERS)," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 3, pp. 516–524, 2004.

[27] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Royal Stat. Soc., Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[28] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused Lasso," *J. Royal Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 67, no. 1, pp. 91–108, 2004.

[29] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.

[30] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial Mathematics, 1990.

[31] N. Weiner, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. Cambridge, MA: MIT Press, 1949.

[32] H. T. Wu, P. Flandrin, and I. Daubechies, "One or two frequencies? The synchrosqueezing answers," *Adv. Adaptive Data Anal.*, vol. 1, no. 2, p. 29–39, 2011.

[33] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2005.

[34] H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE Trans. Signal Processing*, vol. 59, no. 5, pp. 2002–2016, May 2011.

[SP]