Angelosante and Giannakis *EURASIP Journal on Advances in Signal Processing* 2012, **2012**:70
http://asp.eurasipjournals.com/content/2012/1/70

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Group lassoing change-points in piecewise-constant AR processes

Daniele Angelosante[1*] and Georgios B Giannakis[2]

Abstract

Regularizing the least-squares criterion with the total number of coefficient changes, it is possible to estimate time-varying (TV) autoregressive (AR) models with piecewise-constant coefficients. Such models emerge in various applications including speech segmentation, biomedical signal processing, and geophysics. To cope with the inherent lack of continuity and the high computational burden when dealing with high-dimensional data sets, this article introduces a convex regularization approach enabling efficient and continuous estimation of TV-AR models. To this end, the problem is cast as a sparse regression one with grouped variables, and is solved by resorting to the group least-absolute shrinkage and selection operator (Lasso). The fresh look advocated here permeates benefits from advances in variable selection and compressive sampling to signal segmentation. An efficient block-coordinate descent algorithm is developed to implement the novel segmentation method. Issues regarding regularization and uniqueness of the solution are also discussed. Finally, an alternative segmentation technique is introduced to improve the detection of change instants. Numerical tests using synthetic and real data corroborate the merits of the developed segmentation techniques in identifying piecewise-constant TV-AR models.

1. Introduction

Autoregressive (AR) models have been the workhorse for parametric spectral estimation 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 [1], Chap. 3]. These are among the main reasons why AR models have been widely adopted in various applications as diverse as speech modeling [2-5], electroencephalogram (EEG) signal analysis [6], and geophysics [7]. While AR modeling of stationary random processes is well appreciated, a number of signals encountered in real life are non-stationary. This justifies the growing interest toward non-stationary signal analysis and time-varying (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 the changes in the electrical activity of neurons. If the TV-AR coefficient trajectories can be well approximated by superimposing a small number of basis sequences, non-stationary modeling reduces to estimating via, e.g., least-squares (LS), the basis expansion coefficients [7]. On the other hand, it has been well-documented that *piecewise-constant* AR systems excited by white Gaussian noise are capable of modeling real-world signals such as speech and EEG [6-8]. Piecewise-constant AR models constitute a subset of TV-AR models wherein AR coefficients change *abruptly*. In this case, basis expansion techniques fall short in estimating the change points [8].

Exploiting the piecewise constancy of TV-AR models, several methods are available to detect the changing instants of the AR coefficients, and thus facilitate what is often referred to as *signal segmentation*. The literature on signal segmentation is large since the topic is of interest in signal processing, applied statistics, and several other branches of science and engineering. Recent advances can be mainly divided in two categories. The first class adopts regularized LS criteria in order to impose piecewise-constant AR coefficients. To avoid "oversegmentation," the LS cost is typically regularized with the total number of changes [6]. The resulting estimator can be implemented via dynamic programming (DP), which incurs computational burden that scales quadratically with the signal dimension. For large data sets, such as those considered in speech processing, this burden refrains practitioners from applying DP to segmentation, and heuristics are pursued instead based on the generalized likelihood ratio test

* Correspondence: daniele.angelosante@ch.abb.com
[1] Asea Brown Boveri (ABB) Corporate Research Center, Baden, CH 5405, Switzerland
Full list of author information is available at the end of the article

(GLRT), or, approximations of the maximum likelihood approach [2], [[7], p. 401], [9]. The second class of methods relies on Bayesian inference and Markov Chain Monte Carlo (MCMC) methods [3-5]. A distinct advantage of this class is that model order selection can be performed automatically, and a variable model order can be chosen per segment. However, Bayesian techniques are known to require large computational resources.

The algorithm for change detection of piecewise-constant AR models developed in this article belongs to the first class of methods, and its first novelty consists in developing a new regularization function which encourages piecewise-constant TV-AR coefficients while being *convex* and *continuous*; hence, it can afford efficient convex optimization solvers. To this end, it is shown that the segmentation problem can be recast as a *sparse regression* problem. The regularization function in [6] is then relaxed with its tightest convex approximation. It turns out that the resultant change detector is a modification of the group least-absolute shrinkage and selection operator (Lasso) [10].

With the emphasis placed on large data sets, a candidate algorithm for implementing the developed change detector is a block-coordinate descent iteration, which is provably convergent to the group Lasso solution. Surprisingly, it turns out that each iteration of the block-coordinate descent can be implemented at complexity that scales linearly with the signal dimension, thus encouraging its application to large data sets. Regularization tuning and uniqueness of the group Lasso solution are also discussed.

The second novelty of the present study is an alternative change-point retrieval algorithm based on the smoothly-clipped absolute deviation (SCAD) regularization. The associated non-convex problem is tackled by resorting to a local linear approximation (LLA), which yields iterated weighted group Lasso minimization problems that can be solved via block-coordinate descent. Numerical tests using synthetic and real (speech and sound) data are performed to corroborate the capability of the developed algorithms to identify piecewise-constant TV-AR models.

The remainder of the article is structured as follows. Section 2 deals with piecewise-constant TV-AR model estimation preliminaries. In Section 3, the problem at hand is recast as a sparse linear regression, and the novel group Lasso approach is introduced. An efficient block-coordinate descent algorithm is developed in Section 4, while tuning issues and uniqueness of the group Lasso solution are addressed in Section 5. Section 6 introduces a non-convex segmentation method based on the SCAD regularization to enhance the sparsity of the solution, which translates to retrieving more precisely the change instants. Numerical tests are presented in Section 7, and concluding remarks are summarized in Section 8. The Appendix is devoted to technical proofs. *Notation:* Column vectors (matrices) are denoted using lower-case (upper-case) boldface letters; calligraphic letters are reserved for sets; $(\cdot)^T$ stands for transposition, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2; $\otimes$ denotes the Kronecker product; 0_L is the L-dimensional column vector with all zeros, and $\mathbf{I}_L$ is the L-dimensional identity matrix. The ℓ_p norm of $\mathrm{x} := [x_1, \ldots, x_L]^T \in \mathbb{R}^L$ is defined as $\|\mathrm{x}\|_p := \left(\sum_{l=1}^{L} |x_l|^p\right)^{\frac{1}{p}}$.

2. Preliminaries and problem statement

Let $\{y_n\}_{n=-L}^{N}$ denotes the realization of an Lth order TV-AR process obeying the discrete-time input-output relationship

$$y_n = \sum_{l=1}^{L} a_{l,n} y_{n-l} + v_n, \quad n = 0, 1, \ldots, N \tag{1}$$

where v_n denotes the zero-mean white input noise at time n with variance $\sigma^2 := \mathbb{E}\left[v_n^2\right] < +\infty$, and $a_{\ell,n}$ is the ℓth TV-AR coefficient at time n. With $\mathrm{h}_n := [y_{n-1}, y_{n-2}, \ldots, y_{n-L}]^T \in \mathbb{R}^L$ and $\mathrm{a}_n := [a_{1,n}, a_{2,n}, \ldots, a_{L,n}]^T \in \mathbb{R}^L$, (1) can be rewritten as

$$y_n = \mathbf{h}_n^T \mathrm{a}_n + v_n, \quad n = 0, 1, \ldots, N. \tag{2}$$

Suppose that *abrupt changes* in the spectrum of $\{y_n\}$ occur, due to *piecewise-constant* changes of $\mathrm{a_n}$; that is,

$$\mathrm{a}_n = a_k, \quad n_k \leq n \leq n_{k+1} - 1 \tag{3}$$

for k = 0,1,..., 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, n_0 = 0 and n_{K+1} - 1 = N.

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

2.1. Optimum segmentation of TV-AR processes

Modeling real world signals using AR processes is well motivated because, for a given continuous spectral density $S(f)$, it is possible to find an AR process (of high enough order) whose spectral density is arbitrarily close to $S(f)$ [11], p. 130]. On the other hand, depending on the underlying non-stationary phenomena, variations of the AR coefficients can be either slow or abrupt. The problem stated in Section 2 is often referred to as *signal segmentation*, and emerges in numerous applications ranging from speech processing [2-5] to EEG signal analysis [6]. Regularized LS has been the workhorse approach for analyzing this kind of non-stationary processes [12-15]. Denoting with μ a positive tuning constant, a Schwarz-like regularization is typically adopted to estimate jointly the change points and the AR coefficients, i.e.,

$$\{\hat{\mathrm{a}}_n\}_{n=0}^N := \arg\min_{\{\mathrm{a}_n\}_{n=0}^N} \left[\frac{1}{2}\sum_{n=0}^N (y_n - \mathrm{h}_n^T \mathrm{a}_n)^2 + \mu \sum_{n=1}^N \delta_{0_L}(\mathrm{a}_n - \mathrm{a}_{n-1})\right] \quad (4)$$

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

$$\delta_{0_L}(\mathrm{a}) := \begin{cases} 0, \text{ if } \mathrm{a} = 0_L \\ 1, \text{ otherwise.} \end{cases} \quad (5)$$

The non-convex regularization term $\sum_{n=1}^N \delta_{0_L}(\mathrm{a}_n - \mathrm{a}_{n-1})$ not only captures the total number of changes, but also encourages piecewise-constant $\{\hat{\mathrm{a}}_n\}_{n=0}^N$. Clearly, the larger the μ, the smaller the total number of changes. The estimator in (4) is optimal in the maximum a posteriori (MAP) sense when the change occurrences are modeled as Bernoulli random variables, and $v_n \sim \mathcal{N}(0, \sigma^2)$[6]. In some problems, the total number of changes is known, and the following constrained version of (4) is adopted instead:

$$\begin{aligned} \{\hat{\mathrm{a}}_n\}_{n=0}^N &= \arg\min_{\{\mathrm{a}_n\}_{n=0}^N} \sum_{n=0}^N (y_n - \mathrm{h}_n^T \mathrm{a}_n)^2 \\ \text{s.t.} \quad & \sum_{n=1}^N \delta_{0_L}(\mathrm{a}_n - \mathrm{a}_{n-1}) = K. \end{aligned} \quad (6)$$

From a practical point of view, the minimization in (4) or (6) 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 (4) [6,16]. Despite the fact that DP approaches solve (4) in polynomial time, the computational complexity is quadratic in N, which limits their applicability to signal segmentation in practice. In typical applications, N can be very large (up to several thousands), and even quadratic complexity cannot be afforded. On the other hand, when applied to real data, the performance of the estimator in (4) is not satisfactory [17].

To overcome these limitations of (4), heuristic approaches based on the GLRT are used in real world applications [[7], p. 401], [9,18,19]. However, GLRT-based change detectors are sensitive to modeling errors, and require fine tuning of the associated detection thresholds.

In what follows, a convex relaxation of the cost in (4) is advocated based on recent advances in sparse linear regression and compressive sampling. To this end, (4) is first reformulated to a sparse regression problem with non-convex regularization that is successively relaxed through its tightest convex approximation. The consequent optimization rule will yield sparse vector estimators which result in surprisingly accurate retrieval of change-points. Those are obtained by an efficient block-coordinate descent iteration that incurs only linear computational burden and memory storage. Unlike (4), based on well-established results in statistics, it will be further argued that the resultant TV-AR model estimates are a continuous function of the data.

3. Sparse linear regression and group lassoing

Let $\mathrm{y} := [y_0, y_1, \ldots, y_N]^T \in \mathbb{R}^{N+1}$ denotes the observation vector, $\mathrm{a} := \left[\mathrm{a}_0^T, \mathrm{a}_1^T, \ldots, \mathrm{a}_N^T\right]^T \in \mathbb{R}^{(N+1)^L}$, $\mathrm{m}_n := [\underbrace{0_L^T, \ldots, 0_L^T}_{n}, \mathrm{h}_n^T, \underbrace{0_L^T, \ldots, 0_L^T}_{N-n}]^T \in \mathbb{R}^{(N+1)L}$ for n = 0,1,..., N, and $\mathrm{M} := [\mathrm{m}_0, \mathrm{m}_1, \ldots, \mathrm{m}_N]^T \in \mathbb{R}^{N+1\times(N+1)L}$, such that

$$\sum_{n=0}^N (y_n - \mathbf{h}_n^T \mathrm{a}_n)^2 = \|\mathrm{y} - \mathrm{Ma}\|_2^2. \quad (7)$$

Define the "difference" vector $\mathrm{d}_n \in \mathbb{R}^L$ as

$$\mathrm{d}_n = \begin{cases} \mathrm{a}_n, & \text{if } n = 0 \\ \mathrm{a}_n - \mathrm{a}_{n-1}, & \text{otherwise} \end{cases} \quad (8)$$

and $\mathrm{d} := \left[\mathrm{d}_0^T, \mathrm{d}_1^T, \ldots, \mathrm{d}_N^T\right]^T \in \mathbb{R}^{(N+1)L}$. Observe that d_n = 0_L for n >0 if and only if there is no change in the TV-AR coefficients between time instants n - 1 and n. Clearly, it is possible to recover $\{\mathrm{a}_n\}_{n=0}^N$ from $\{\mathrm{d}_n\}_{n=0}^N$ since

$$\mathrm{a}_n = \sum_{n'}^{n} \mathrm{d}_{n'}. \quad (9)$$

Let $\mathrm{T} \in \mathbb{R}^{N+1\times N+1}$ denotes a lower triangular matrix with all nonzero entries equal to one and $\mathrm{X} := \mathrm{M}(\mathrm{T} \otimes \mathrm{I}_L) \in \mathbb{R}^{(N+1)\times(N+1)L}$, having the following structure:

$$\mathrm{X} = \begin{bmatrix} \mathrm{h}_0^T & 0_L^T & \cdots & \cdots & 0_L^T \\ \mathrm{h}_1^T & \mathrm{h}_1^T & 0_L^T & \cdots & 0_L^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{h}_{N-1}^T & \mathrm{h}_{N-1}^T & \mathrm{h}_{N-1}^T & \mathrm{h}_{N-1}^T & 0_N^T \\ \mathrm{h}_N^T & \mathrm{h}_N^T & \mathrm{h}_N^T & \mathrm{h}_N^T & \mathrm{h}_N^T \end{bmatrix}. \quad (10)$$

Since a = $(\mathbf{T} \otimes \mathbf{I}_L)$d, an equivalent formulation of (4) in terms of $\{\mathrm{d}_n\}_{n=0}^N$ can be given as [cf. (7)]

$$\{\hat{\mathrm{d}}_n\}_{n=0}^N := \arg\min_{\{\mathrm{d}_n\}_{n=0}^N} \left[\frac{1}{2}\|\mathrm{y} - \mathrm{Xd}\|_2^2 + \mu \sum_{n=1}^N \delta_{0_L}(\mathrm{d}_n)\right] (11)$$

What makes the formulation in (11) attractive but also challenging is the non-convex and discontinuous Schwarz-like regularization term. The latter “pushes” most of the $\{\mathrm{d}_n\}_{n=1}^N$ vectors toward 0_L, while d_0 is not penalized. As a consequence, the vector $\hat{\mathrm{d}} := \left[\hat{\mathrm{d}}_0^T, \hat{\mathrm{d}}_1^T, \ldots, \hat{\mathrm{d}}_N^T\right]^T$ is *group* sparse, and the non-zero group indexes correspond to the change instants of the TV-AR coefficients. Recently, a *convex* model selector with grouped variables was put forth by [10], and successfully applied to biostatistics and compressive sampling [20]. It generalizes the (non-grouped) least-absolute shrinkage and selection operator (Lasso) [21] 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 Schwarz-like regularization in (11) with its *tightest* convex approximation.

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

$$\{\widehat{\mathrm{d}}_n\}_{n=0}^N = \arg\min_{\{\mathrm{d}_n\}_{n=0}^N} \left[\frac{1}{2}\left\|\mathrm{y}-\mathrm{X}\mathrm{d}\right\|_2^2 + \lambda\sum_{n=1}^N \left\|\mathrm{d}_n\right\|_2\right] 1(12)$$

where λ is a positive tuning parameter. It is known that the group Lasso regularization encourages group sparsity; that is, $\hat{\mathrm{d}}_n = 0_L$ for most $n > 0$ [10]. Again, the larger the λ, the sparser the $\hat{\mathrm{d}}$.

The role of the regularization term in (12) is illustrated next through a simple example. Select $L = 2$ for simplicity, and let $d := [d_1, d_2]^T$. Consider the family of penalties $\|d\|_2^p = (d_1^2 + d_2^2)^{\frac{p}{2}}$, where $0 < p \leq 2$. Figure 1 depicts the penalties $\|d\|_2^p$ for $p = 2,1,0.5$, and 0.1. Clearly, $\|d\|_2^p$ is convex for $1 \leq p \leq 2$. On the other hand, $\|d\|_2^p$ is non-convex for $0 < p < 1$ but it comes close *to* $\delta_{0_L}(d)(d)$ as p approaches 0. This demonstrates that $\left\|\mathrm{d}_n\right\|_2$ is the tightest convex approximation of $\delta_{0_L}(\mathrm{d}_n)$ (d_n). Furthermore, $\left\|\mathrm{d}_n\right\|_2$ is non-differentiable at $\mathrm{d}_n = 0_L$, 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 next, with computational complexity per iteration that scales linearly with N. But first two remarks are in order.

Remark 1. Different from Schwarz-like regularization, Figure 1 illustrates that the group Lasso one grows unbounded. This makes the resultant estimator biased. Nevertheless, unlike the Schwarz-like regularization, the group Lasso one is continuous which renders the resulting estimator more stable when applied to real data; see also [10,22]. A continuous regularization function that reduces the bias of the group Lasso will be discussed in Section 6.

Remark 2. Convex relaxation for detecting changes in the mean of non-stationary processes was recently mentioned in [12], and analyzed in [17]. For the mean-change problem, the tightest convex approximation of the Schwarz-like regularized LS is provided by the Lasso, which can afford efficient solvers such as the least-angle regression (LARS) algorithm [23]. However, for the group Lasso cost proposed here to catch changes in TV-AR models, an exact LARS-like solver is not available [10];

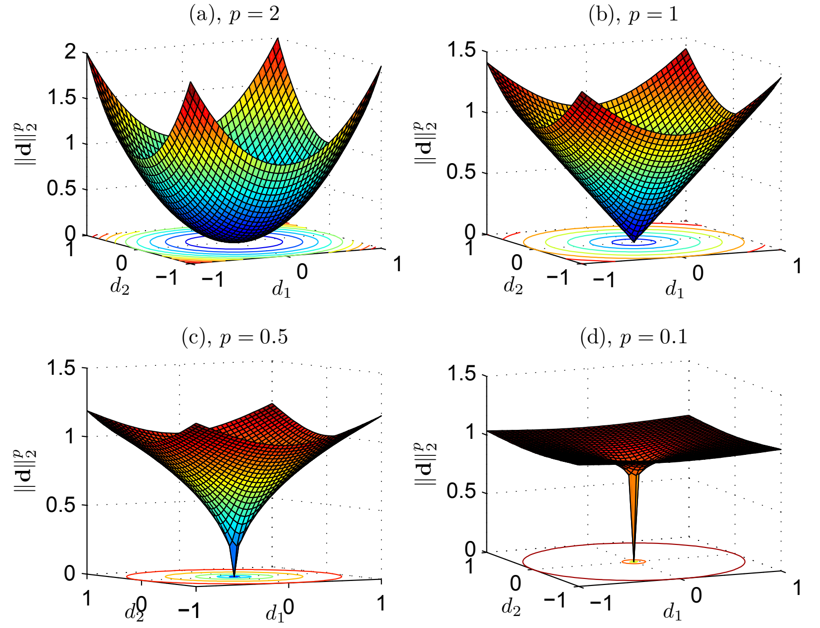

Figure 1 Regularization family $\|d\|_2^p$ for L = 2 and (a)p = 2, (b)p = 1, (c)p = 0.5, and (d)p = 0.1.

thus, the pursuit of efficient algorithms for solving (12) is well motivated. This is the theme of the ensuing section.

4. Block-coordinate descent solver

The crux of block-coordinate descent is to iterate minimizations of the function of interest over a group of variables, while keeping the rest fixed. Consider the objective function

$$J(\mathrm{d}) := \frac{1}{2}\left\|\mathrm{y} - \mathrm{X}\mathrm{d}\right\|_2^2 + \lambda \sum_{n=1}^{N} \left\|\mathrm{d}_n\right\|_2 \qquad (13)$$

and let $\mathrm{d}^{(i-1)} := [\mathrm{d}_0^{(i-1)^T}, \mathrm{d}_1^{(i-1)^T}, \ldots, \mathrm{d}_N^{(i-1)^T}]^T$ denotes the provisional solution at iteration *i*-1. The *n*th step of the *i*th block-coordinate descent iteration entails minimization of *J*(*d*) *only* with respect to d_n, while retaining the provisional estimates at iteration *i*-1, namely $\{\mathrm{d}_{n'}^{(i-1)}\}_{n'=n+1}^{N}$, and the newly updated blocks at iterations *i*, namely $\{\mathrm{d}_{n'}^{(i)}\}_{n'=0}^{n-1}$. Thus, block-coordinate descent at the *n*th step of the *i*th iteration yields

$$\mathrm{d}_n^{(i)} = \arg\min_{\mathrm{d}_n} J\left(\left[\mathrm{d}_0^{(i)}, \ldots, \mathrm{d}_{n-1}^{(i)}, \mathrm{d}_n, \mathrm{d}_{n+1}^{(i-1)}, \ldots, \mathrm{d}_N^{(i-1)}\right]\right) \qquad (14)$$

for *n* = 0,1,..., *N*, and *i* >0. Skipping constant terms, *J* (d) in (13) can be rewritten as

$$J(\mathrm{d}) = \frac{1}{2}\mathrm{d}^T\mathrm{X}^T\mathrm{X}\mathrm{d} - \mathrm{d}^T\mathrm{X}^T\mathrm{y} + \lambda\sum_{n=1}^{N}\left\|\mathrm{d}_n\right\|_2 = \frac{1}{2}\mathrm{d}^T\mathrm{R}\mathrm{d} - \mathrm{d}^T\mathrm{r} + \lambda\sum_{n=1}^{N}\left\|\mathrm{d}_n\right\|_2 \qquad (15)$$

where $\mathrm{R} := \mathrm{X}^T\mathrm{X}$, and $\mathrm{r} := \mathrm{X}^T\mathrm{y}$. Upon defining $\mathrm{R}_{n:n'} := \sum_{m=n}^{n'} \mathrm{h}_m\mathrm{h}_m^T$ and $\mathrm{r}_{n:n'} := \sum_{m=n}^{n'} \mathrm{h}_m y_m$ for *n'* ≥ *n* , it holds that

$$\mathrm{R} = \begin{bmatrix} \mathrm{R}_{0:N} & \mathrm{R}_{1:N} & \cdots & \mathrm{R}_{N-1:N} & \mathrm{R}_{N:N} \\ \mathrm{R}_{1:N} & \mathrm{R}_{1:N} & \cdots & \mathrm{R}_{N-1:N} & \mathrm{R}_{N:N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{R}_{N-1:N} & \mathrm{R}_{N-1:N} & \cdots & \mathrm{R}_{N-1:N} & \mathrm{R}_{N:N} \\ \mathrm{R}_{N:N} & \mathrm{R}_{N:N} & \cdots & \mathrm{R}_{N:N} & \mathrm{R}_{N:N} \end{bmatrix} \qquad (16)$$

and

$$\mathrm{r} = \begin{bmatrix} \mathrm{r}_{0:N} \\ \mathrm{r}_{1:N} \\ \vdots \\ \mathrm{r}_{N-1:N} \\ \mathrm{r}_{N:N} \end{bmatrix}. \qquad (17)$$

While for *n* = 0 (14) reduces to an LS problem, for *n* >0, omitting again irrelevant terms, it can be rewritten as

$$\mathrm{d}_n^{(i)} = \arg\min_{\mathrm{d}_n \in \mathbb{R}^L}\left[\frac{1}{2}\mathrm{d}_n^T\mathrm{R}_{n:N}\mathrm{d}_n + \mathrm{d}_n^T\mathrm{g}_n^{(i)} + \lambda\left\|\mathrm{d}_n\right\|_2\right] \qquad (18)$$

with

$$\mathrm{g}_n^{(i)} := \mathrm{R}_{n:N}\left(\sum_{n'=0}^{n-1}\mathrm{d}_{n'}^{(i)}\right) + \sum_{n'=n+1}^{N}\mathrm{R}_{n':N}\mathrm{d}_{n'}^{(i=1)} - \mathrm{r}_{n:N}. \qquad (19)$$

The problem in (18) is a convex second-order cone program (SOCP). Typically, $L \ll N$ and (18) can be solved with fast optimization solvers based on interior point methods [24], at worst-case complexity $\mathcal{O}(L^{3.5})$. Recently, it has been shown that the solution of (18) can be obtained as a function of the solution of the following *scalar* problem

$$\gamma_n^{(i)} := \arg\min_{\gamma \geq 0}\left[\gamma\left(1 - \frac{1}{2}\mathrm{g}_n^{(i)^T}\left(\gamma\mathrm{R}_{n:N} + \frac{\lambda^2}{2}\mathrm{I}_L\right)^{-1}\mathrm{g}_n^{(i)}\right)\right] \qquad (20)$$

whose solution is given by [25]

$$\gamma_n^{(i)} = \begin{cases} 0, & \text{if } \left\|\mathrm{g}_n^{(i)}\right\|_2 \leq \lambda \\ \gamma > 0 : \quad \left\|\frac{\lambda}{2}\left(\gamma\mathrm{R}_{n:N} + \frac{\lambda^2}{2}\mathbf{I}_L\right)^{-1}\mathrm{g}_n^{(i)}\right\|_2^2 = 1, & \text{otherwise.} \end{cases} \qquad (21)$$

Finally, $\mathrm{d}_n^{(i)}$ in (18) can be obtained from $\gamma_n^{(i)}$ in (21) as

$$\mathrm{d}_n^{(i)} = \begin{cases} 0_L, & \text{if } \left\|\mathrm{g}_n^{(i)}\right\|_2 \leq \lambda \\ -\gamma_n^{(i)}\left(\gamma_n^{(i)}\mathrm{R}_{n:N} + \frac{\lambda^2}{2}\mathbf{I}_L\right)^{-1}\mathrm{g}_n^{(i)}, & \text{otherwise.} \end{cases} \qquad (22)$$

Notice that if $\left\|\mathrm{g}_n^{(i)}\right\|_2 \leq \lambda$, the solution of (18) is $\mathrm{d}_n^{(i)} = 0_L$. Since it is expected that the solution of (12) is sparse, solving (18) is trivial most of the time. If $\left\|\mathrm{g}_n^{(i)}\right\|_2 > \lambda$, $\mathrm{d}_n^{(i)}$ can be obtained via interior point methods or by (numerically) solving the scalar equation in (21), which admits fast solvers via, e.g., Newton-Raphson iterations, as in [25].

Despite the fact that block-coordinate descent is typically adopted for large-size sparse linear regression, what makes it particularly appealing for catching change-points is the fact that the vector $\mathrm{g}_n^{(i)}$ can be updated recursively in *n* due to the special structure of **R** in (16). Upon defining

$$\mathrm{c}_n^{(i)} := \sum_{n'=0}^{n=1}\mathrm{d}_{n'}^{(i)} \qquad (23)$$

$$\mathrm{s}_n^{(i)} := \sum_{n'=n+1}^{N}\mathrm{R}_{n':N}\mathrm{d}_{n'}^{(i-1)} \qquad (24)$$

it follows from (19) that

$$\mathrm{g}_n^{(i)} = \mathrm{R}_{n:N}\mathrm{c}_n^{(i)} + \mathrm{s}_n^{(i)} - \mathrm{r}_{n:N} \qquad (25)$$

which shows that evaluating $g_n^{(i)}$ requires the vectors $c_n^{(i)}$ and $s_n^{(i)}$ Given $\{d_n^{(i-1)}\}_{n=0}^{N}$ from the (i - 1)st iteration, and initializing $c_n^{(i)}$ and $s_n^{(i)}$ at n = 0 as $c_0^{(i)} = 0_L$ and $s_0^{(i)} = \sum_{n=1}^{N} R_{n:N} d_n^{(i-1)}$, it is possible to recursively evaluate $c_n^{(i)}$ and $s_n^{(i)}$ given $c_{n-1}^{(i)}$, $s_{n-1}^{(i)}$ and $d_{n-1}^{(i)}$ from step n-1 for n >0 as

$$c_n^{(i)} = c_{n-1}^{(i)} + d_{n-1}^{(i)} \tag{26}$$

$$s_n^{(i)} = s_{n-1}^{(i)} - R_{n:N} d_n^{(i-1)}. \tag{27}$$

The block-coordinate descent algorithm is summarized in Algorithm 1. Interestingly, matrix $X \in \mathbb{R}^{(N+1)\times(N+1)L}$ in (12) does not have to be stored since only $\{R_{n:N}\}_{n=0}^{N}$ and $\{r_{n:N}\}_{n=0}^{N}$ suffice to implement Algorithm 1. Thus, the memory storage and complexity to perform one block-coordinate descent iteration grow linearly with N. This attribute renders the block-coordinate descent appealing especially for large-size problems where DP approaches tend to be too expensive.

Regarding convergence, the ensuing assertion is a direct consequence of the results in [26].

Proposition 1. *The iterates* $d^{(i)} := \left[d_0^{(i)^T}, d_1^{(i)^T}, \ldots, d_N^{(i)^T}\right]^T$ *obtained by* Algorithm 1 *converge to the global minimum of* (12)*; that is,* $\lim_{i\to\infty} d^{(i)} = \hat{d}$·

Block-coordinate descent will also be the basic building block for solving the non-convex problem introduced in Section 6 to improve the retrieval of change-points. But first, it is useful to consider two issues of the group Lasso change detector for TV-AR models.

```
Given $\{R_{n:N}, r_{n:N}\}_{n=0}^{N}$
Initialize with $d_n^{(0)} = 0_L$ for $n$ = 1, ... , $N$
for $i > 0$ do
  for $n$ = 0,1,..., $N$ do
    if $n$ = 0 then
      $c_0^{(i)} = 0_L$
      $s_0^{(i)} = \sum_{n=1}^{N} R_{n:N} d_{n-1}^{(i-1)}$
      $g_0^{(i)} = s_0^{(i)} - r_{0:N}$
      $d_0^{(i)} = -R_{0:N}^{-1} g_0^{(i)}$
    else
      $c_n^{(i)} = c_{n-1}^{(i)} + d_{n-1}^{(i)}$
      $s_n^{(i)} = s_{n-1}^{(i)} - R_{n:N} d_n^{(i-1)}$
      $g_n^{(i)} = R_{n:N} c_n^{(i)} + s_n^{(i)} - r_{n:N}$
      if $\left\|g_n^{(i)}\right\|_2 \leq \lambda$then
        $d_n^{(i)} = 0_L$
      else
        $d_n^{(i)} = \arg\min_{d_n \in \mathbb{R}^L} \left[\frac{1}{2} d_n^T R_{n:N} d_n + d_n^T g_n^{(i)} + \lambda \left\|d_n\right\|_2\right]$
```

Algorithm 1: **Block-coordinate descent algorithm**

5. Regularization and uniqueness issues

Performance of model selection with grouped variables via group Lasso and related approaches has been analyzed in [10,20,27], while asymptotic analysis has been pursued in [28,29]. In this section, a couple of issues are investigated regarding the group Lasso cost function, and the uniqueness of its minimum.

5.1. Tuning the regularization parameter

Selection of λ is a critical issue since larger λ's promote sparser solutions, which translate to fewer changes in the TV-AR spectrum. However, larger λ's increase the estimator bias as well. If the number of changes present are known a priori by other means, or, if a certain level of segmentation can be afforded, λ can be tuned accordingly by 'trial and error,' or by cross-validation. But in general, analytic methods to automatically choose the "best" value of λ are not available. In essence, selecting the regularization parameters is more a matter of engineering art, rather than systematic science.

In this section, heuristic but useful guidelines are provided to choose λ based on rigorously established lower bounds of this parameter. Given $X \in \mathbb{R}^{N+1\times(N+1)L}$ in (10), define $X_n \in \mathbb{R}^{N+1\times L}, n = 0, 1, \ldots, N$ such that X = [X_0, X_1,..., X_N]. To bound λ, we will rely on the following result; see Appendix 1 for the proof.

Proposition 2. *If* X_0 *has full column rank, then* $\hat{d} = [d_{0,c}^{T}, 0_L^T, \ldots, 0_L^T]^T$ *with* $d_{0,c} := (\mathbf{X}_0^T \mathbf{X}_0)^{-1} \mathbf{X}_0^T y$, *if and only if* $\lambda \geq \lambda^* : \max_{n=1,\ldots,N} \left\|\mathbf{X}_n^T(\mathbf{X}_0 d_{0,c} - y)\right\|_2$

If λ exceeds a threshold, which is specified by the regression matrix and the observations, Proposition 2 asserts that $\hat{d}_0 = d_{0,c}$ and $\hat{d}_n = 0_L$ for n = 1,..., N. This, along with (9), implies that $\hat{a}_n = d_{0,c}$ for n = 0, 1,..., N; that is, no change occurs in the coefficients of the TV-AR process. To avoid this trivial (change-free) solution, the guideline provided by Proposition 2 is that λ must be chosen strictly less than λ*. Our extensive simulations suggest that setting λ equal to a small percentage of λ*, say 5-20%, results in satisfactory estimates.

5.2. Uniqueness of the sparse solution

Uniqueness of sparse linear regression with non-grouped variables has been studied in [30-32]. Next, uniqueness issues in recovering sparse vectors with group-variables are explored by exploiting the deterministic structure of the regression matrix in (12). The cost function in (12) is not strictly convex since X is a fat matrix, and the regularization term is not strictly convex; see also Figure 1. On the other hand, the block-coordinate descent algorithm developed in Section 4 is guaranteed to converge to a global minimum. In the following, a criterion is introduced to check a posteriori whether the obtained solution is unique for a given group-sparsity level.

Traditionally, the support of a sparse vector is defined as the set of indexes corresponding to the non-zero entries. In the group-sparsity framework herein, a different definition of support is required. Indeed, the vector of interest here, namely $\mathrm{d} = [\mathrm{d}_0^T, \mathrm{d}_1^T, \ldots, \mathrm{d}_N^T]^T$ comprises $N + 1$ groups of L-dimensional variables. Since the term d0 is not penalized in (12), $\hat{\mathrm{d}}_0 \neq 0_L$ almost surely. Define the *group support* (g-support) of $\hat{\mathrm{d}}$ to be the set containing the indexes relative to the nonzero group of variables; that is, $\mathrm{g}\text{-}\mathrm{supp}(\hat{\mathrm{d}}) := \{n \in \{1, \ldots, N\} : \hat{\mathrm{d}}_n \neq 0_L\}$ In the following, when $\mathcal{G} := \{s_1, \ldots, s_{|\mathcal{G}|}\} \subset \{1, \ldots, N\}$ denotes the g-support of d, the set $\mathcal{G}$ is assumed ordered; i.e., $s_j < s_k$ for each $j < k$.

The following lemma establishes a property of the matrix X in (10); see Appendix 2 for the proof.

Lemma 1. *If any L out of* $N + 1$ *vectors* $\{\mathrm{h}_n\}_{n=0}^N$ *are linearly independent, for any g-support* $\mathcal{G} = \{s_1, \ldots, s_{|\mathcal{G}|}\}$ *such that* $(|\mathcal{G}| + 1)L \leq N + 1, s_1 \geq L, s_{|\mathcal{G}|} \leq N - L + 1,$ *and* $|s_j - s_k| \geq L$ *for each* $j \neq k$*, the matrix* $\mathrm{X}_{\mathcal{G}} := [\mathrm{X}_0, \mathrm{X}_{s1,\ldots,}\mathrm{X}_{s_{|\mathcal{G}|}}] \in \mathbb{R}^{N+1\times(|\mathcal{G}|+1)L}$ *has full column rank.*

Lemma 1 asserts full column rank of the submatrix $\mathrm{X}_{\mathcal{G}}$, if it is formed by the columns of X corresponding to the non-zero indexes of any sparse vector whose g-support is sufficiently small, and the non-zero groups are sufficiently distant from each other.

Next, Lemma 1 is exploited to establish an interesting property for the solutions of (12); see Appendix 3 for the proof.

Proposition 3. *If any L out of N+1 vectors* $\{\mathrm{h}_n\}_{n=0}^N$ *are linearly independent, for any g-support* $\mathcal{G} = \{s_1, \ldots, s_{|\mathcal{G}|}\} \subset \{1, \ldots, N\}$ *such that* $(|\mathcal{G}| + 1)L \leq N + 1, s_1 \geq L, s_{|\mathcal{G}|} \leq N - L + 1,$ *and* $|s_j - s_k| \geq L$ *for each* $j \neq k$*, there exists at most one solution of* (12) *g-supported in* $\mathcal{G}$.

Proposition 3 ensures that if $\hat{\mathrm{d}}$ is g-supported in $\mathcal{G}$, and is sufficiently sparse with non-zero groups sufficiently far apart, then $\hat{\mathrm{d}}$ is the *only* solution of (12) g-supported in $\mathcal{G}$.

Remark 3. Analysis of the group Lasso and its modifications has revealed that its performance can be close to the Schwarz-regularized LS either when the regression matrix is sufficiently block incoherent, or, when the block restricted isometry property holds [10,27]. If the regression matrix can be chosen by the designer, and it is randomly drawn from selected distributions (e.g., Gaussian or Bernoulli), these analyzes provide useful connections between problems (11) and (12). In the problem at hand however, the regression matrix is fixed, and its blocks $[\mathrm{X}_0, \mathrm{X}_1, \ldots, \mathrm{X}_N]$ are highly correlated. In this case, the relationship between the solutions of (11) and (12) is much less understood, and constitutes an interesting future research direction.

6. Continuity, bias and the group SCAD

As already pointed out, convex relaxation of the Schwarz-like regularization was developed in [12,17] for the mean-change problem using the (non-grouped) Lasso. Numerical results in [17] reveal that the Lasso tends to detect a "cloud" of small change points around an actual change. Post-processing via DP to select a few of the estimated change instants was proposed in [17]. Moreover, due to the bias introduced by the Lasso, once the change points are obtained, another step is required to re-estimate the mean within a segment. In the following, a novel change detector is developed based on recent advances in model selection via non-convex regularization. The resulting estimator reduces the bias of group Lasso and can afford a convergent optimization solver. The corresponding algorithm is based on iterative instantiations of weighted group Lasso, which is capable of enhancing the sparsity of the solution [33,34], and thus improving the precision of the detected change points.

Attributes of a "good" regularization function are delineated in [22], and three properties are identified to this end:

- *Unbiasedness*. The estimator has to be unbiased when the true unknown parameter has large amplitude.
- *Sparsity*. The estimator has to set small-amplitude coefficients to zero to reduce the number of variables.
- *Continuity*. The estimator has to be *continuous* in the data to avoid instability when estimating (non-) zero variables.

To appreciate these properties, a simple setting is presented next. Consider estimating a scalar parameter, call it d, in additive noise v, based on the scalar observation $y = d + v$. In its general form, the regularized LS approach yields

$$\hat{d} = \arg\min_{d} \left[\frac{1}{2}(y - d)^2 + p_\lambda(|d|) \right] \tag{28}$$

where $p_\lambda(|\mathrm{d}|)$ is the regularization function. The Lasso regularization is $p_\lambda^{\mathrm{Lasso}}(|d|) = \lambda\,|d|$ while the Schwarz-like regularization is

$$p_{\lambda}^{\text{Schwarz}}(|d|)=\begin{cases}0, & \text{if } d=0\\ \lambda, & \text{otherwise.}\end{cases}\tag{29}$$

In both cases, the estimate $\hat{d}$ can be found in closed form, and the dependence of $\hat{d}$ on y is given as

$$\hat{d}^{\text{Lasso}}=\begin{cases}0, & \text{if } |y|<\lambda\\ \text{sign}(y)\left(|y|-\lambda\right), & \text{otherwise}\end{cases}\tag{30}$$

$$\hat{d}^{\text{Schwarz}}=\begin{cases}0, & \text{if } |y|<\sqrt{2\lambda}\\ y, & \text{otherwise.}\end{cases}\tag{31}$$

The non-linear estimation rules in (30) and (31) are depicted in Figure 2a,b, respectively, for λ = 2. Observe that both regularization functions effect sparsity, since coefficients with small amplitude are set to 0. The Schwarz-like regularization yields unbiased estimates, but the solution is *not* continuous with respect to y. Hence, small variations of y or λ may result in large variations of $\hat{d}$ (this happens when one is uncertain whether to set the coefficient to 0 or not). On the other hand, the Lasso regularization possesses continuity but the estimates are biased, because in addition to small, large-amplitude coefficients are "shrunk" too.

To overcome the limitations of these regularization functions, the following smoothly clipped absolute deviation (SCAD) regularization can be used with a >2 [22]

$$p_{\lambda}^{\text{SCAD}}(|d|)=\begin{cases}\lambda|d|, & \text{if } |d|\le\lambda\\ -\dfrac{d^2-2|d|a\lambda+\lambda^2}{2(a-1)}, & \text{if } \lambda<|d|\le a\lambda\\ \dfrac{\lambda^2}{2}(a+1), & |d|>a\lambda\end{cases}\tag{32}$$

to obtain estimates given by

$$\hat{d}^{\text{SCAD}}=\begin{cases}0 & \text{if } |y|\le\lambda\\ \text{sign}(y)(|y|-\lambda), & \text{if } \lambda<|d|\le 2\lambda\\ \dfrac{(a-1)y-\text{sign}(y)a\lambda}{a-2}, & \text{if } 2\lambda<|y|\le a\lambda\\ y, & |y|>a\lambda.\end{cases}\tag{33}$$

The data dependence of $\hat{d}^{\text{SCAD}}$ is depicted in Figure 2c for λ = 2 and a = 3.7. Observe that the SCAD enjoys the three aforementioned attributes of a desirable regularization function.

Motivated by this scalar example, we propose as an alternative to (12), the following group SCAD approach for catching change-points:

$$\{\hat{\mathrm{d}}_n\}_{n=0}^{N}=\arg\min_{\{\mathrm{d}_n\}_{n=0}^{N}}\left[\frac{1}{2}\|\mathrm{y}-\mathrm{X}\mathrm{d}\|_2^2+\sum_{n=1}^{N}p_{\lambda}^{\text{SCAD}}(\|\mathrm{d}_n\|_2)\right].\tag{34}$$

The problem in (34) is non-convex and its exact minimization is challenging due to the presence of local minima. Nevertheless, it is possible to generalize the iterated local linear approximation (LLA) of [34] in order to ensure converge to a stationary point of the cost in (34). Let $p_{\lambda}^{\text{SCAD}'}(d)$ denote the derivative of $p_{\lambda}^{\text{SCAD}}(d)$for d≥0; that is,

$$p_{\lambda}^{\text{SCAD}'}(d):=\begin{cases}\lambda, & \text{if } d\le\lambda\\ \dfrac{a\lambda-d}{(a-1)}, & \text{if } \lambda<d\le a\lambda\\ 0, & d>a\lambda.\end{cases}\tag{35}$$

The idea behind the LLA is to approximate $p_{\lambda}^{\text{SCAD}'}(|d|)$ with its linear expansion around d_o; that is

$$p_{\lambda}^{\text{SCAD}}(|d|)\approx p_{\lambda}^{\text{SCAD}}(|d_o|)+p_{\lambda}^{\text{SCAD}'}(|d_o|)(|d|-|d_o|).\tag{36}$$

Given a provisional estimate of $\{\mathrm{d}_n\}_{n=0}^{N}$ at iteration j - 1, namely $\left\{\hat{\mathrm{d}}_n^{[j-1]}\right\}_{n=0}^{N}$, the iterated LLA of (34) is

$$\left\{\hat{\mathrm{d}}_n^{[j]}\right\}_{n=0}^{N}=\arg\min_{\{\mathrm{d}_n\}_{n=0}^{N}}\left[\frac{1}{2}\|\mathrm{y}-\mathrm{X}\mathrm{d}\|_2^2+\sum_{n=1}^{N}p_{\lambda}^{\text{SCAD}'}\left(\left\|\hat{\mathrm{d}}_n^{[j-1]}\right\|_2\right)\|\mathrm{d}_n\|_2\right]\tag{37}$$

for j = 1,..., J. Since $p_{\lambda}^{\text{SCAD}'}\left(\left\|\hat{\mathrm{d}}_n^{[j-1]}\right\|_2\right)$ are non-negative constants, the cost in (37) is convex, and can be minimized using the block-coordinate descent algorithm of Section 4. The role of the weights $p_{\lambda}^{\text{SCAD}'}\left(\left\|\hat{\mathrm{d}}_n^{[j-1]}\right\|_2\right)$

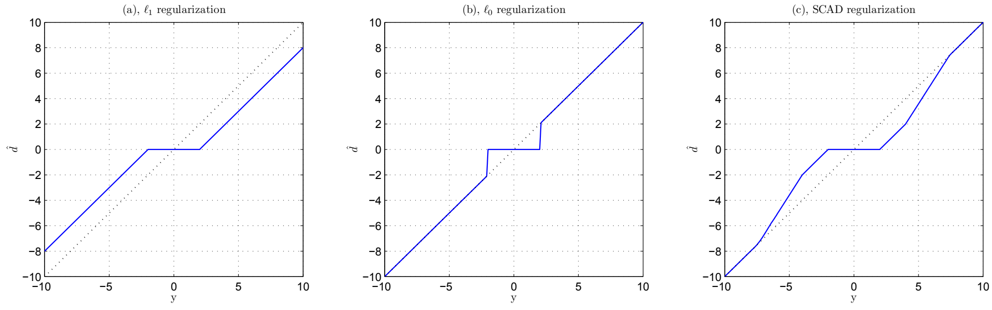

Figure 2 Regularized LS estimation rules in the scalar case: (a) (group) Lasso, **(b)** Schwarz-type regularization, and **(c)** SCAD.

is to avoid penalizing terms that, most likely, are non-zero. Function $p_{\lambda}^{\text{SCAD}'}(|d|)$ is depicted in Figure 3 for λ = 2, a = 3.7, and $d \geq 0$. It is clearly a decreasing function of its argument. More precisely, if $0 \leq d \leq \lambda$, then $p_{\lambda}^{\text{SCAD}'}(d) = \lambda$ hence, the regularization parameter is as large as the group Lasso. If $\lambda \leq d \leq a\lambda$, the regularization parameter is linearly decreasing until $d \geq a\lambda$, where the regularization parameter is zero. Thus, the expression in (37) represents an iterated *weighted* group Lasso. Furthermore, if initialized with $\hat{\mathrm{d}}^{[0]} = 0_{(N+1)L} = 0_{(N+1)L}$, the first iteration of (37) corresponds to the (unweighted) group Lasso, and few iterations of the type in (37) are required for convergence.

Remark 4. In principle, one could also apply a block-coordinate descent iteration to minimize (34) directly. The resulting iterates converge to a local minimum of (34) that depends on the starting points; see also [26]. In general, it is impossible to assess properties of this solution. Instead, the solution of the LLA in (37) has provable merits in estimating the true support of sparse signals [34].

Remark 5. Recently, greedy algorithms such as the matching pursuit and the orthogonal matching pursuit have been shown to approach the performance of (11) and (12) when the regression matrix is sufficiently block incoherent, or, when the block restricted isometry property holds [27]. In the problem at hand, wherein the regression matrix exhibits correlation among blocks, simulated tests have shown that greedy algorithms suffer from severe error propagation. In fact, a cloud of change points is typically declared around a true change point. For these reasons, greedy algorithms will not be considered hereafter.

Remark 6. As already mentioned in the Introduction, one advantage of Bayesian techniques is that automatic model selection can be performed on a per-segment basis at the price of increasing computational cost. The convex relaxation developed in Section 3 can be modified to perform model selection too. Indeed, setting L to a prescribed upper bound on the model order, one may further regularize the cost in (12) to impose sparse AR coefficient vectors; that is,

$$\{\hat{\mathrm{a}}_n\}_{n=0}^{N} = \arg \min_{\{\mathrm{a}_n\}_{n=0}^{N}} \left[\frac{1}{2} \|\mathrm{y} - \mathrm{Ma}\|_2^2 + \lambda \sum_{n=1}^{N} \|\mathrm{a}_n - \mathrm{a}_{n-1}\|_2 + \gamma \|\mathrm{a}\|_1 \right]. \quad (38)$$

The cost function in (38) is convex and effects a piece-wise-constant and sparse TV-AR model. It is different from model order selection criteria in the sense that the selected non-zero AR coefficients do not necessarily have to be consecutive. A challenge associated with the optimization in (38) is that block-coordinate descent algorithms do not converge, since the differentiable part is not separable group-wise [26]. The cost function in (38) resembles the fused Lasso developed in [35]. Efficient algorithms exploiting this link, and the structure of the problem at hand are currently under investigation.

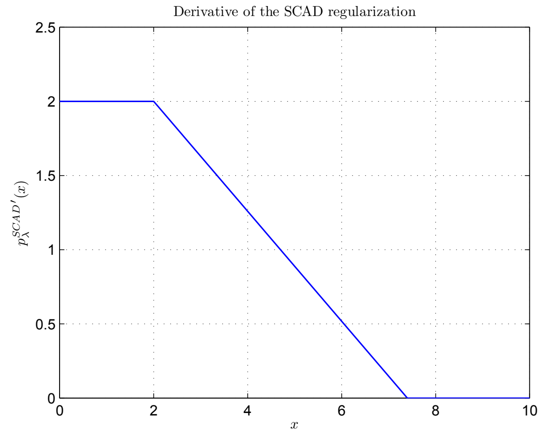

Figure 3 Derivative of the SCAD function.

7. Simulated tests

The merits of the novel approaches to catching change-points in TV-AR processes are assessed via numerical simulations using synthetic and real data.

7.1. Synthetic data

The signal of interest here is a realization of a TV-AR process with $N + 1 = 500$, order $L = 4$, $v_n \sim \mathcal{N}(0, \sigma^2)$, and $\sigma^2 = 10^{-2}$, exhibiting $K = 2$ abrupt changes in the spectrum at time $n_1 = 100$, and $n_2 = 350$. The AR model during the first segment ($n \in [0,99]$) has coefficients a_0 = [-0.8000, -0.1500, 0.1940, -0.0280]^T, while in the second segment ($n \in [100,349]$) the AR coefficients are a_1 = [0.1200, 0.0245, -0.2787, -0.0693]^T. In the third segment ($n \in [350, 499]$), a_0 is in act. Figure 4 shows the true signal along with the group Lasso, group SCAD, and DP-based estimates of the TV-AR coefficients obtained via (12), (37), and (6) with $K = 2$, respectively. The regularization parameter λ of the group Lasso and group SCAD was selected to return 2 change points. The block-coordinate descent algorithm of Section 4 was used to solve (12) up to a maximum of 10^3 iterations or when $\frac{\|\mathrm{d}^{(i)} - \mathrm{d}^{(i-1)}\|_2^2}{\|\mathrm{d}^{(i)}\|_2^2} \leq 10^{-8}$ The same stopping rule is used for each iteration in (37), and $J = 5$ outer iterations are run (the same stopping rules are adopted henceforth). Figure 5 depicts the variation of $\left\|\hat{\mathrm{d}}_n\right\|_2$ across time. Clearly, the instants where $\left\|\hat{\mathrm{d}}_n\right\|_2 > 0$ represents the change points retrieved. While the DP has detected change-points at $\hat{n}_1 = 105$ and $\hat{n}_2 = 349$ the group Lasso

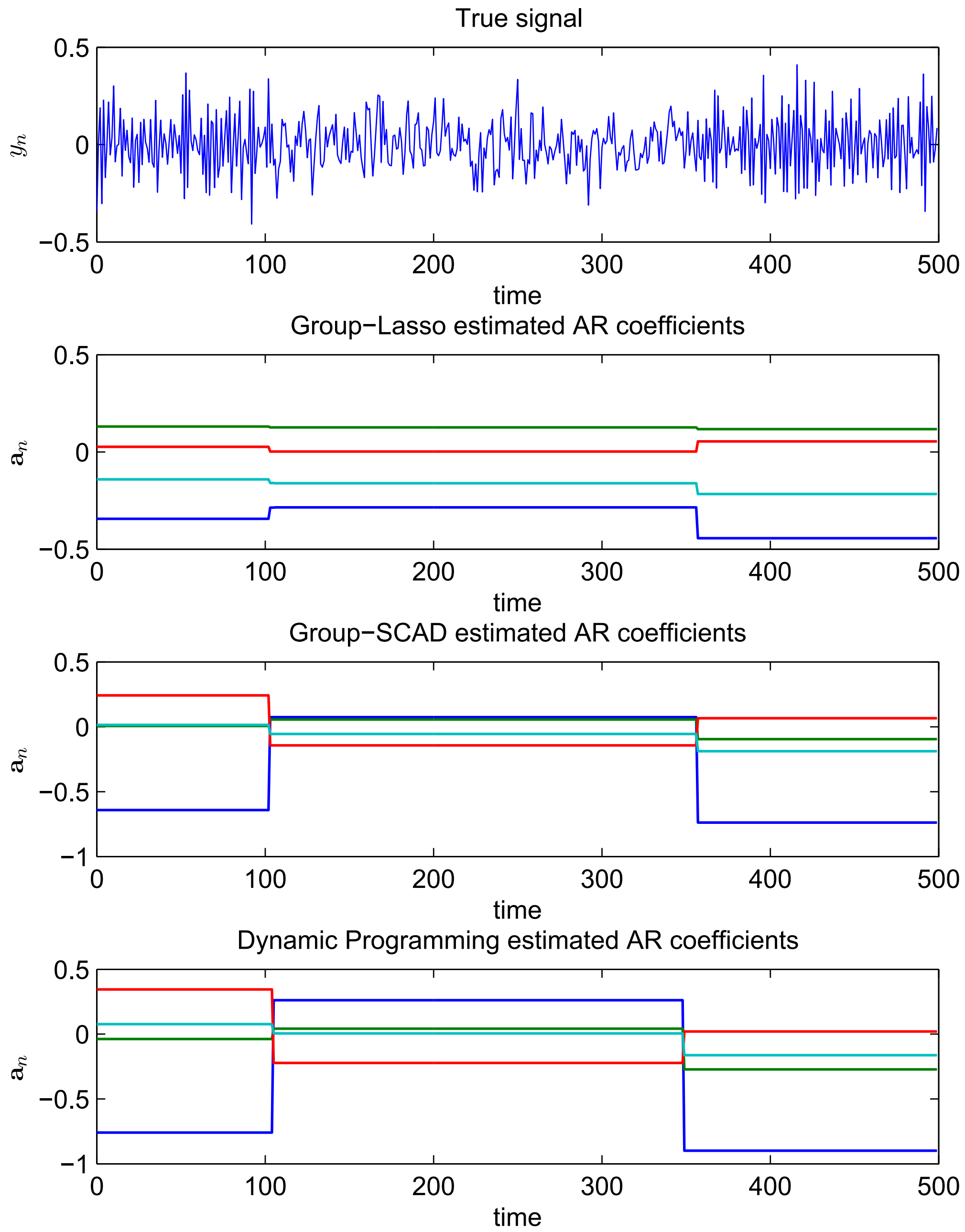

Figure 4 Synthetic data. From top to bottom: True signal, estimated of the TV-AR coefficients by group Lasso, group SCAD, and DP, respectively.

and the group SCAD have detected changes at $\hat{n}_1 = 103$ and $\hat{n}_2 = 356$ confirming that the results returned by the DP programming, by the group Lasso, and by group SCAD are comparable. Since $\{y_n\}$ was generated as in (1), any L out of $N + 1$ vectors $\{\mathrm{h}\ _n\}_{n=0}^N$ are linearly independent almost surely, which means that Proposition 3 can be invoked to ensure uniqueness of the group Lasso TV-AR model estimate. Notice from Figure 4 and Figure 5 that the $\left\|\hat{\mathrm{d}}_n\right\|_2$s estimated by the group Lasso are much smaller than the true ones due to the bias, which results in poor estimates of the AR coefficients. However, those estimated by the group SCAD are closer to the true one.

7.2. Real data: piano sound

Next, a piano sound of 0.5 s comprising three monochromatic notes is sampled at 8 kHz to obtain $N + L + 1 = 4000$ samples. The signal is depicted in Figure 6. A TV-AR model with $L = 1$ is adopted, and λ is selected

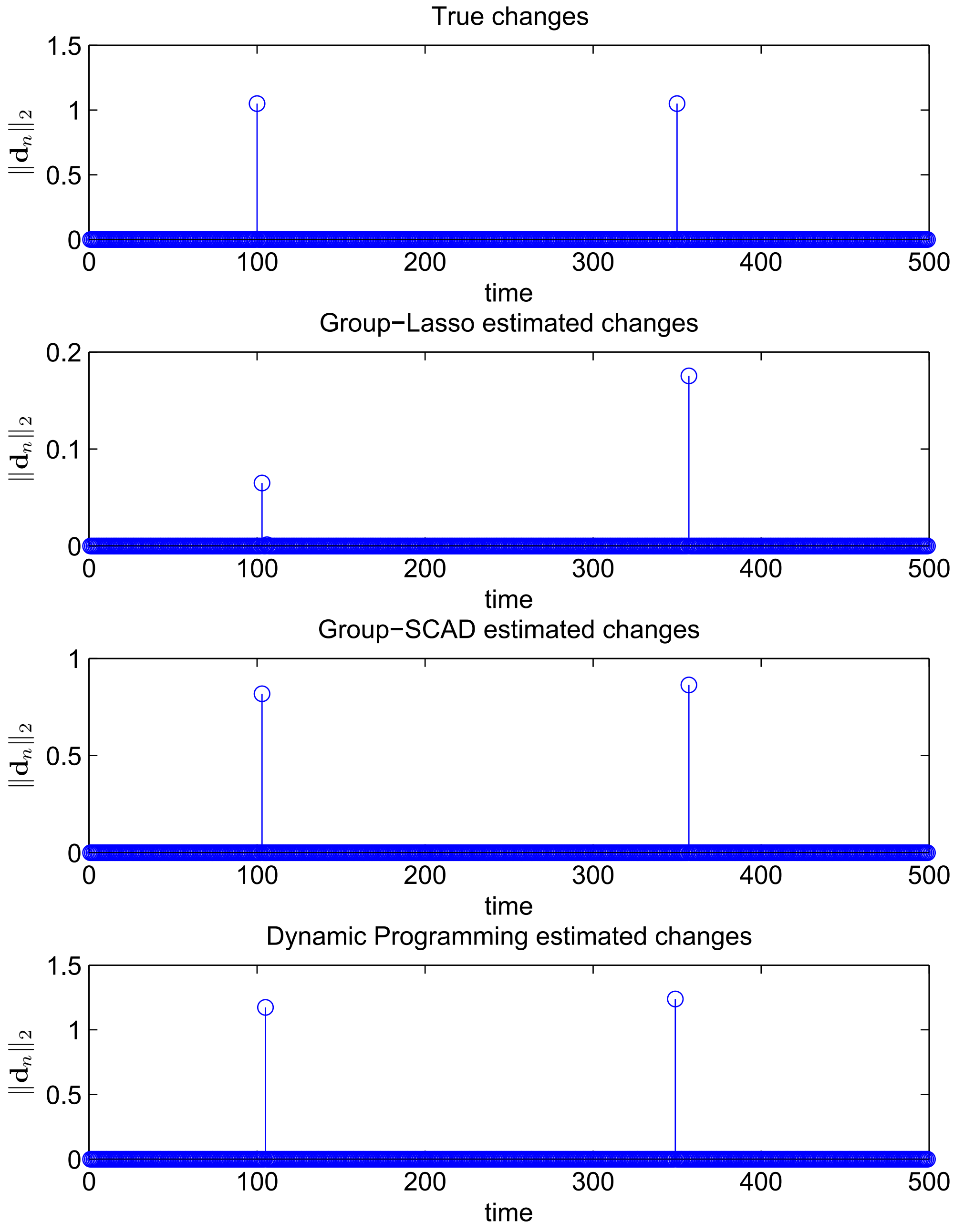

Figure 5 Synthetic data. From top to bottom: True TV-AR coefficients change, changes detected by group Lasso, group SCAD, and DP, respectively.

to be λ*/10. Figure 7 depicts the TV-AR coefficients estimated by the group Lasso and the group SCAD. The estimated $\left\|\hat{\mathrm{d}}_n\right\|_2$ over time is displayed in Figure 8. Observe that the group Lasso captures the correct changes at time instants $\hat{n}_1$ = 1632 and $\hat{n}_2$ = 3155 along with a small *false* change at time instant n_f = 2770. Instead, the group SCAD retrieves two changes at time instants $\hat{n}_1$ = 1632 and $\hat{n}_2$ = 3155 Notice also that group SCAD exhibits a reduced bias in the estimated $\left\|\hat{\mathrm{d}}_n\right\|_2$ which results in better estimates of the TV-AR coefficients.

7.3. Real data: speech

Here, a speech signal of 0.5 s is sampled at 8 kHz, to obtain $N + L + 1 = 4000$ samples. The resultant time series depicted in Figure 9 comprises a descent

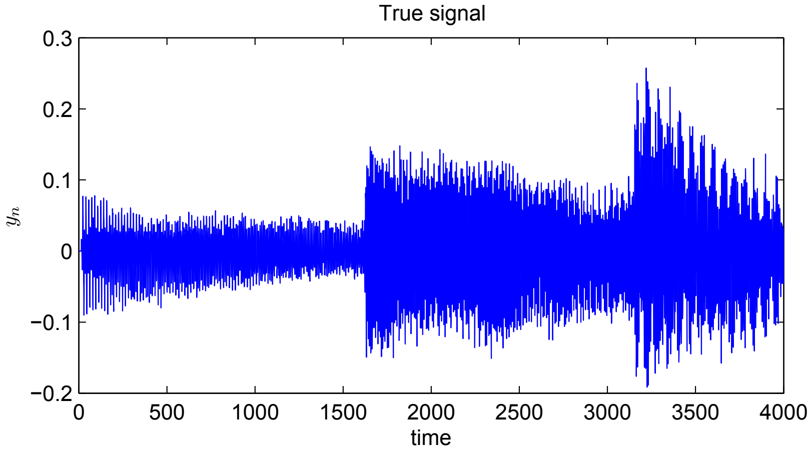

Figure 6 Piano sound comprising three monochromatic notes.

diphthong /ai/ followed by an /o/, pronounced by another party. A TV-AR model with $L = 8$ is adopted, and λ is selected to be λ*/10. The change of vocoid in the diphthong occurs approximately at time instant n_1 = 1500, while the /o/ occurs approximately at n_2 = 3000. Figure 10 shows the TV-AR coefficients estimated by the group Lasso and group SCAD. In agreement with [17], 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. Figure 11 depicts $\|\mathrm{d}_n\|_2$ over time. The group SCAD detects four segments with change points at $\hat{n}_1$ = 1065, $\hat{n}_3$ = 2993 and $\hat{n}_3$ = 2993. 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 forth to the /o/. Observe that the group Lasso exhibits peaks around the actual change instants, and requires post-processing via either DP as advocated in [17], or, by simply peak picking.

Further tests are performed on a sampled speech that has been widely adopted for TV-AR speech segmentation [3-5,7]. According to [7, p. 401], this signal belongs to a database designed by the French National Agency for Telecommunications (CNET) for testing and evaluating speech recognition algorithms. It consists of a noisy French speech recorded in a car at sampling

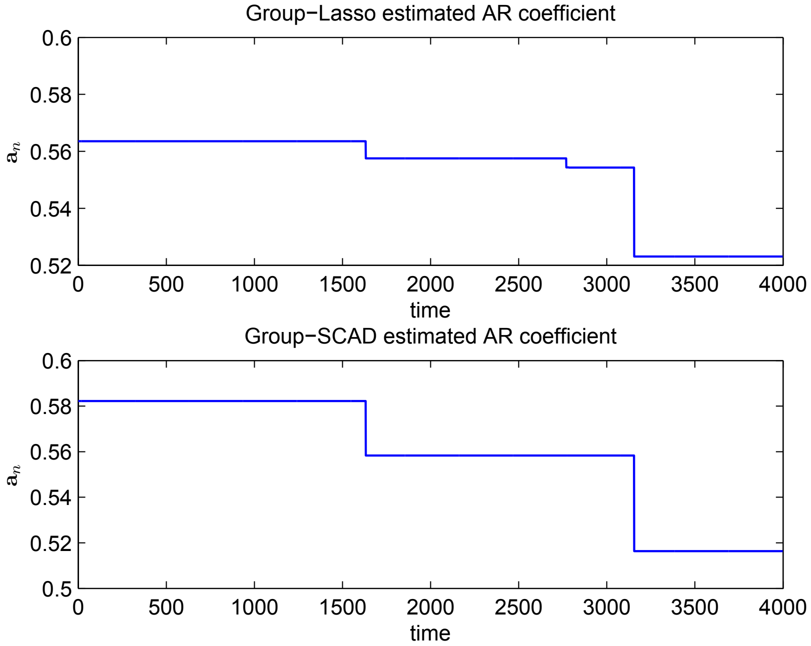

Figure 7 Piano sound. Estimated of the TV-AR coefficients by group Lasso (top), and group SCAD (bottom).

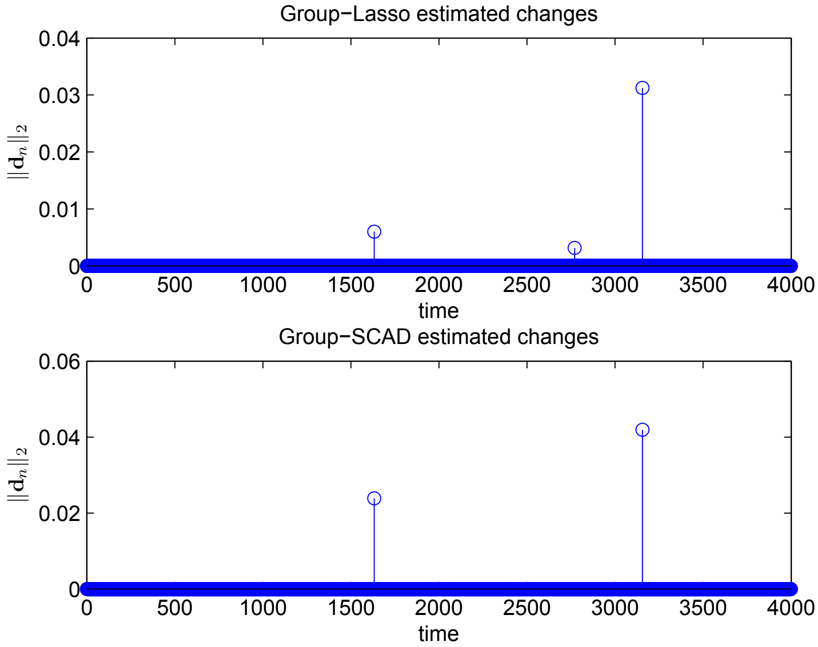

Figure 8 Piano sound. Changes detected by group Lasso (top), and group SCAD (bottom).

frequency 8 kHz, prefiltered by a highpass filter with cutoff frequency equal to 150 Hz, and quantized with 16 bits per sample. The time series is shown in Figure 12 along with the changes caught by the group SCAD change detector, and the GLR algorithm of [18] whose results are reported in [3-5] for $L = 2$. The group SCAD is tuned to return the same number of changes as the GLR. The detected change instants are listed in Table 1.

The first change detected by the GLR is at sample 445, while this change is not detected by the group SCAD. Interestingly, it is reported in [3] that this change is not relevant for segmentation purposes, and this fact is apparent by inspection of the true signal. Both algorithms have detected changes around samples 1750, 2100, 2800, and 3650. The group SCAD successfully removed the false change detected by the GLR at sample 3400. By inspecting the original signal, this change does not seem to be relevant. The group SCAD had detected a change instant around sample 1300 unlike the GLR. This change is also detected by advanced Bayesian techniques reported in [3-5]. The group SCAD has detected a change at sample 779, while the GLR at sample 645. Indeed, inspection of the original signal suggests that the detection of the GLR is preferable. Surprisingly, the group SCAD, unlike the GLR and the Bayesian techniques of [3-5], has detected a change at sample 2359. Observing the original signal around this point, there is a clear amplitude modulation

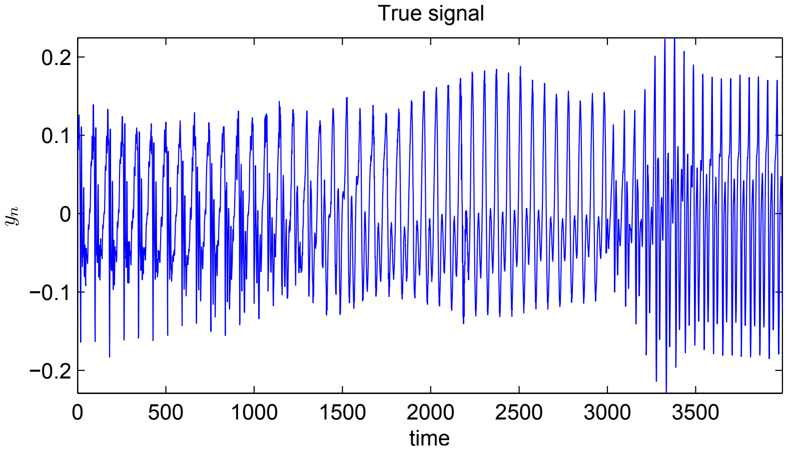

Figure 9 Speech signal: /ai//o/.

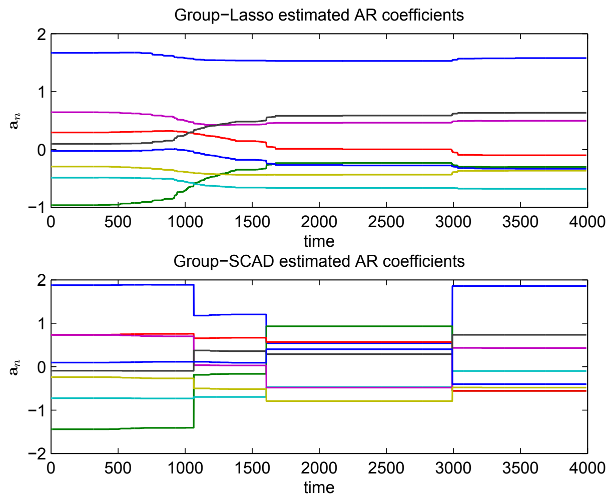

Figure 10 Speech signal. Estimated of the TV-AR coefficients by group Lasso (top), and group SCAD (bottom).

that may cause a change in the TV-AR coefficients, which existing algorithms have passed undetected.

A way to univocally compare the two algorithms is via the segmented prediction error (SPE). Assuming that K changes have been detected at instants $\{\hat{n}_k\}_{k=1}^{K}$, let $\hat{a}_k$ denotes the LS estimates of the AR model in the kth segment, i.e., $\hat{a}_k = \arg\min_{\mathrm{a}\,\in\mathbb{R}^L} \sum_{n=\hat{n}_k}^{\hat{n}_{k+1}-1} (y_n - \mathrm{h}_{\;n}\mathrm{a}\;)^2$. The SPE is defined as SPE := $\sum_{k=0}^{K} \sum_{n=\hat{n}}^{\hat{n}_{k+1}-1} (y_n - h_n\hat{a}_k)^2$, and represents the error in approximating the original signal $\{y_n\}$ with a TV-AR model exhibiting abrupt changes at instants $\{\hat{n}_k\}_{k=1}^{K}$. The GLR segmentation entails $\mathrm{SPE}_{\mathrm{glr}} = 0.2638$, while $\mathrm{SPE}_{\mathrm{g\text{-}scad}} = 0.2578$ for the group SCAD. Clearly, the group SCAD based segmentation seems preferable to that of the GLR algorithm.

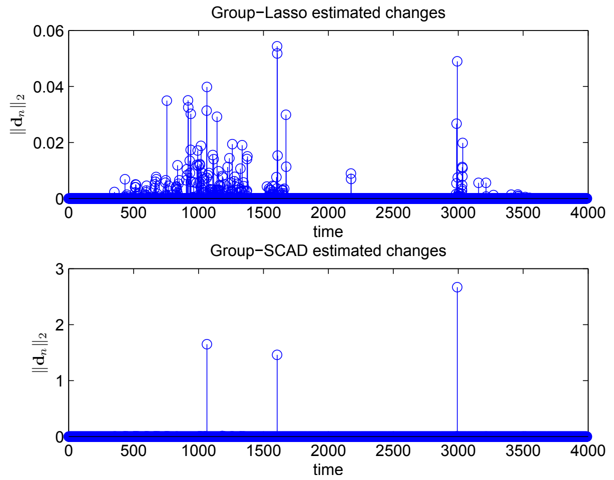

Figure 11 Speech signal. Changes detected by group Lasso (top), and group SCAD (bottom).

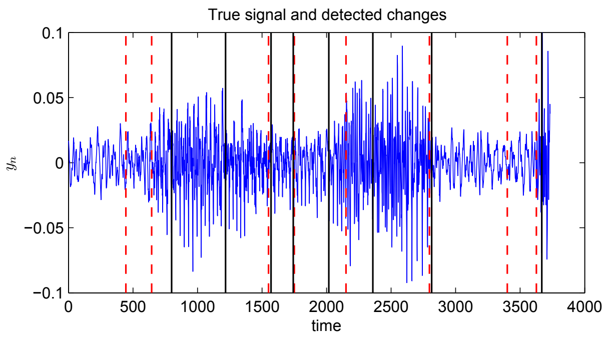

Figure 12 French speech signal and changes detected by the group SCAD (dot) and GLR (dash) change TV-AR detectors for$L = 2$.

Finally, since general analysis of the convergence rate for the block-coordinate descent algorithm is not available, a simulated test assessing its converge speed is performed. Figure 13 depicts the normalized step size amplitude variations, namely $\|\mathrm{d}^{\,(i)} - \mathrm{d}^{\,(i-1)}\|_2^2/(\|\mathrm{d}^{\,(i)}\|_2^2)$, across the iteration index (i) for the first LLA of the group SCAD (which amounts to a group Lasso since $\hat{\mathrm{d}}^{[0]} = 0_{(N+1)L}$ for the speech signal in Figure 12. Observe that after a few hundred iterations, the normalized step amplitude size drops below 10^{-6}. For practical purposes, the solution at this stage might be acceptable. Progressively, the speed of convergence slows down since the algorithm has to decide whether some components are truly zero or have small amplitudes, and large jumps in $\|\mathrm{d}^{\,(i)} - \mathrm{d}^{\,(i-1)}\|_2^2/(\|\mathrm{d}^{\,(i)}\|_2^2)$ correspond to iteration indexes where the small components are set to zero. Once the correct vector support is determined, the speed of converge is fast. It is worth pointing out that only 10 min are required to perform 5,000 iterations of the block-coordinate descent (implemented with Matlab) for the problem at hand, whereas advanced Bayesian techniques might require several hours (see e.g., [3]).

8. Concluding remarks

Novel estimators were developed in this article for identification of piecewise-constant TV-AR models by exploiting recent advances in variable selection and compressive sampling. While traditional techniques consist in regularizing a LS criterion with the total number of coefficient changes, the novel estimator relies on a convex regularization function, which resembles the group Lasso and can afford efficient implementation

Table 1 Change instants detected by the group SCAD and GLR algorithm

Group SCAD	799	1217	1570	1742	2017	2359	2814	3668
GLR	445	645	1550	1750	2151	2797	3400	3626

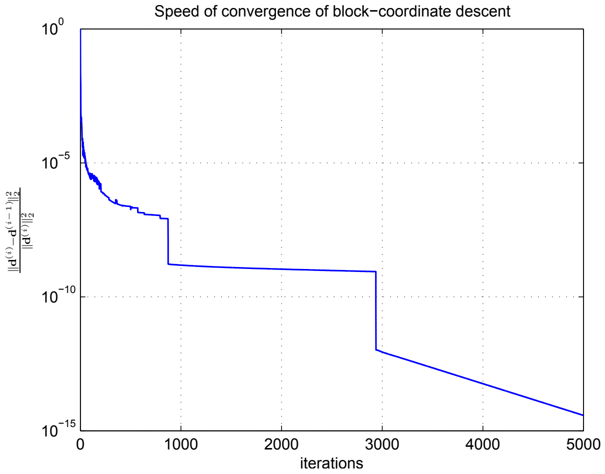

Figure 13 Speed of the block-coordinate descent for the first LLA of the group SCAD.

using block-coordinate descent iterations. The latter incurs computational burden that scales linearly with the number of data samples, thus being particularly attractive for large-size problems. Regularization tuning issues are discussed along with conditions for uniqueness of the estimated piecewise-constant AR model. An alternative group smoothly-clipped absolute deviation regularization is also introduced, and an algorithm based on iterative weighted group Lasso minimizations is developed. Numerical tests using synthetic and real data confirm that the developed algorithms can effectively identify piecewise-constant AR models of large size at manageable complexity, and outperform heuristic alternatives that are based on the GLRT.

Appendix 1: proof of proposition 2

The necessary and sufficient first-order optimality condition for $\hat{\mathrm{d}}$ to be the (unconstrained) minimum of (12), is that the subgradient of J(d) in (13) evaluated at $\hat{\mathrm{d}}$ contains the zero vector [[36], p. 92]; i.e.,

$$\breve{\nabla} J(\hat{\mathrm{d}}) \ni 0_{(N+1)L}. \tag{39}$$

Defining $\mathrm{w} \in \mathbb{R}^{(N+1)L}$ as w := X^T(Xd - y) and $\mathrm{w} = [\mathrm{w}_0^T, \mathrm{w}_1^T, \ldots, \mathrm{w}_N^T]^T$, with $\mathrm{w}_n \in \mathbb{R}^L$ for n = 0,1,..., N, the subgradient of J(d) evaluated at $\hat{\mathrm{d}}$ is given by

$$\breve{\nabla} J(\hat{\mathrm{d}}) = \mathrm{w} + \lambda \mathrm{b} \tag{40}$$

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

$$\mathrm{b}_n := \begin{cases} 0_L, & n = 0 \\ \dfrac{\hat{\mathrm{d}}_n}{\left\|\hat{\mathrm{d}}_n\right\|^2}, n = 1, \ldots, N, \hat{\mathrm{d}}_n \neq 0_L \\ \mathrm{s}_n, & n = 1, \ldots, N. \quad \hat{\mathrm{d}}_n \neq 0_L \end{cases} \tag{41}$$

with $s_n \in \mathbb{R}^L$ such that $\|\mathrm{s}_n\|_2 \leq 1$. Using (40) and (41), Equation (39) translates to the following conditions:

(c1) $\mathrm{w}_0 = 0_l$; and,

$$\text{(c2)}\begin{cases} \mathrm{w}_n + \lambda \dfrac{\hat{\mathrm{d}}_n}{\left\|\hat{\mathrm{d}}_n\right\|_2} = 0_L, \text{ if } \hat{\mathrm{d}}_n \neq 0_L \\ \|\mathrm{w}_n\|_2 \leq \lambda \qquad \text{if } \hat{\mathrm{d}}_n \neq 0_L \end{cases} \text{for } n = 1, \ldots, N.$$

The change-free solution corresponds to having $\hat{\mathrm{d}}_n = 0_L$ for n = 1,..., N. Thus, (c1) implies that $\mathrm{X}_0^T(\mathrm{X}_0\hat{\mathrm{d}}_0 - y) = 0_L$, which is uniquely satisfied by $\hat{\mathrm{d}}_0 = \mathrm{d}_{0,c}$, since X_0 has full column rank. Hence, $\hat{\mathrm{d}}_0 = \mathrm{d}_{0,c}$ and $\hat{\mathrm{d}}_n = 0_L$ for n = 1,..., N hold if and only if (c2) is satisfied, which corresponds to $\|\mathrm{w}_n\|_2 \leq \lambda$ for n = 1,..., N. Since $\mathrm{w}_n = \mathbf{X}_n^T(\mathbf{X}_0\mathrm{d}_{0,c} - \mathrm{y})$, condition (c2) is satisfied if and only if $\lambda \geq \lambda^* := \max_{n=1,\ldots,N}\left\|\mathrm{X}_n^T(\mathrm{X}_0\mathrm{d}_{0,c} - y)\right\|_2$.

Appendix 2: proof of lemma 1

Observe that

$$\begin{aligned}\mathbf{X}_{\mathcal{G}}^T\mathbf{X}_{\mathcal{G}} = {} & \begin{bmatrix} \mathbf{R}_{s_{|\mathcal{G}|}:N} & \mathbf{R}_{s_{|\mathcal{G}|}:N} & \cdots & \mathbf{R}_{s_{|\mathcal{G}|}:N} \\ \mathbf{R}_{s_{|\mathcal{G}|}:N} & \mathbf{R}_{s_{|\mathcal{G}|}:N} & \cdots & \mathbf{R}_{s_{|\mathcal{G}|}:N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{s_{|\mathcal{G}|}:N} & \mathbf{R}_{s_{|\mathcal{G}|}:N} & \cdots & \mathbf{R}_{s_{|\mathcal{G}|}:N} \end{bmatrix} + \begin{bmatrix} \mathbf{R}_{s_{|\mathcal{G}|-1}:s_{|\mathcal{G}|}-1} & \cdots & \mathbf{R}_{s_{|\mathcal{G}|-1}:s_{|\mathcal{G}|}-1} & \mathbf{0}_{L\times L} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{R}_{s_{|\mathcal{G}|-1}:s_{|\mathcal{G}|}-1} & \cdots & \mathbf{R}_{s_{|\mathcal{G}|-1}:s_{|\mathcal{G}|}-1} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} & \cdots & \mathbf{0}_{L\times L} \end{bmatrix} \\ & + \cdots + \begin{bmatrix} \mathbf{R}_{0:s_1-1} & \mathbf{0}_{L\times L} & \cdots & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} & \cdots & \mathbf{0}_{L\times L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} & \cdots & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} & \cdots & \mathbf{0}_{L\times L} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}} \\ \mathbf{h}_{s_{|\mathcal{G}|}} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}} \\ \mathbf{h}_{s_{|\mathcal{G}|}} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}} \\ \mathbf{h}_{s_{|\mathcal{G}|}} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}} \\ \mathbf{h}_{s_{|\mathcal{G}|}} \end{bmatrix}^T + \cdots + \begin{bmatrix} \mathbf{h}_N \\ \mathbf{h}_N \\ \vdots \\ \mathbf{h}_N \\ \mathbf{h}_N \end{bmatrix} \begin{bmatrix} \mathbf{h}_N \\ \mathbf{h}_N \\ \vdots \\ \mathbf{h}_N \\ \mathbf{h}_N \end{bmatrix}^T}_{N-s_{|\mathcal{G}|}+1} \\ & + \underbrace{\begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{0}_L \end{bmatrix} \begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{0}_L \end{bmatrix}^T + + \begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{0}_L \end{bmatrix} \begin{bmatrix} \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \vdots \\ \mathbf{h}_{s_{|\mathcal{G}|}-1} \\ \mathbf{0}_L \end{bmatrix}^T}_{s_{|\mathcal{G}|}-s_{|\mathcal{G}|-1}} + + \underbrace{\begin{bmatrix} \mathbf{h}_0 \\ \mathbf{0}_L \\ \vdots \\ \mathbf{0}_L \\ \mathbf{0}_L \end{bmatrix} \begin{bmatrix} \mathbf{h}_0 \\ \mathbf{0}_L \\ \vdots \\ \mathbf{0}_L \\ \mathbf{0}_L \end{bmatrix}^T + + \begin{bmatrix} \mathbf{h}_{s_1-1} \\ \mathbf{0}_L \\ \vdots \\ \mathbf{0}_L \\ \mathbf{0}_L \end{bmatrix} \begin{bmatrix} \mathbf{h}_{s_1-1} \\ \mathbf{0}_L \\ \vdots \\ \mathbf{0}_L \\ \mathbf{0}_L \end{bmatrix}^T}_{s_1}. \end{aligned} \tag{42}$$

Notice that the first sub-sum in (42) comprises $N - s_{|\mathcal{G}|} + 1$ rank-1 matrices, the last sub-sum comprises s_1 rank-1 matrices, while the gth sub-sum comprises $s_{|\mathcal{G}|-g+2} - s_{|\mathcal{G}|-g+1}$ rank-1 matrices for $g = 2, \ldots, |\mathcal{G}|$. Since $\mathcal{G}$ is such that $s_1 \geq L, s_{|\mathcal{G}|} \leq N - L + 1$, and $|s_j$ - $s_k| \geq L$ for each $j \neq k$, and any L out of N + 1 vectors $\{\mathrm{h}_n\}_{n=0}^N$ are linearly independent, each of the summands in (42) has rank L. Thus, it is possible to find L linearly independent vectors $\{\mathrm{h}_{1,\ell}\}_{\ell=1}^L \subset \mathbb{R}^L$ such that the first sub-sum in (42) equals to $\sum_{\ell=1}^L \tilde{\mathrm{h}}_{1,\ell}\tilde{\mathrm{h}}_{1,\ell}^T$ with $\tilde{\mathrm{h}}_{1,\ell} := [\mathrm{h}_{1,\ell}^T, \ldots, \mathrm{h}_{1,\ell}^T]^T \in \mathbb{R}^{(|\mathcal{G}|+1)L}$. Analogously, it is possible to find L linearly independent vectors $\{\mathrm{h}_{g,\ell}\}_{\ell=1}^L \subset \mathbb{R}^L$ such that the gth sub-sum in (42) can be written as $\sum_{\ell=1}^L \tilde{\mathrm{h}}_{g,\ell}\tilde{\mathrm{h}}_{g,\ell}^T$ with $\tilde{\mathrm{h}}_{g,\ell} := [\underbrace{\mathbf{h}_{g,\ell}^T, \ldots, \mathbf{h}_{g,\ell}^T}_{|\mathcal{G}|-g+2}, \underbrace{\mathbf{0}_L^T, \ldots, \mathbf{0}_L^T}_{g-1}]^T \in \mathbb{R}^{(|\mathcal{G}|+1)L}$ for $g = 2, \ldots, |\mathcal{G}|$. Finally, it is possible to find L linearly independent vectors $\{\mathbf{h}_{|\mathcal{G}|+1,\ell}\}_{\ell=1}^L \subset \mathbb{R}^L$ such that the last sub-sum in (42) can be written as $\sum_{\ell=1}^L \tilde{\mathrm{h}}_{|\mathcal{G}|+1,\ell}\tilde{\mathrm{h}}_{|\mathcal{G}|+1,\ell}^T$ with $\tilde{\mathbf{h}}_{|\mathcal{G}|+1,\ell} := [\mathbf{h}_{|\mathcal{G}|+1,\ell}^T, \mathbf{0}_L^T, \ldots, \mathbf{0}_L^T]^T \in \mathbb{R}^{(|\mathcal{G}|+1)L}$. Thus, $\mathbf{X}_{\mathcal{G}}^{\mathcal{T}}\mathbf{X}_{\mathcal{G}} = \sum_{g=1}^{|\mathcal{G}|+1}\sum_{\ell=1}^L \tilde{\mathbf{h}}_{g,\ell}\tilde{\mathbf{h}}_{g,\ell}^T$, and since $\{\tilde{\mathbf{h}}_{g,\ell}\}$ are ($|\mathcal{G}|$ +

1) linearly independent vectors, $X_{\mathcal{G}}$ has full-column rank.

Appendix 3: proof of proposition 3

Suppose that $\hat{d}$ and $\hat{d}^{'}$ are two solutions of (12) with the same g-support $\mathcal{G}$. Let $X_{\mathcal{G}}$ denotes the matrix obtained by selecting the columns of X relative to the nonzero indexes dictated by the set $\{0\} \cup \mathcal{G}$. The minimization of (13) over the vector having g-support in $\mathcal{G}$ amounts to

$$\hat{u} := \arg \min_{u := [u_0^T, u_0^T, \ldots, u_{\mathcal{G}}^T]^T \in \mathbb{R}^{(|\mathcal{G}|+1)L}} \left[\frac{1}{2} \| y - X_{\mathcal{G}} u \|_2^2 + \lambda \sum_{s=1}^{\mathcal{G}} \| u_s \|_2 \right]. \quad (43)$$

From Lemma 1, $X_{\mathcal{G}}$ has full column rank which implies that $\frac{1}{2} \| y - X_{\mathcal{G}} u \|_2^2$ is strictly convex, and so is $\frac{1}{2} \| y - X_{\mathcal{G}} u \|_2^2 + \lambda \sum_{s=1}^{\mathcal{G}} \| u_s \|_2$. Thus, (43) admits a unique solution.

Since both $\hat{d}$ and $\hat{d}^{'}$ are g-supported in $\mathcal{G}$ by hypothesis, and their restrictions to $\{0\} \cup \mathcal{G}$ are equal to û, it follows readily that $\hat{d} = \hat{d}^{'}$.

Acknowledgements
This work was supported by MURI (AFOSR FA9550-10-1-0567) grant.

Author details
[1]Asea Brown Boveri (ABB) Corporate Research Center, Baden, CH 5405, Switzerland [2]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Competing interests

The authors declare that they have no competing interests.

This paper was originally submitted to the IEEE Transactions on Signal Processing on April 6, 2010, and rejected on June 8, 2010 - date on which we became aware of the independent work by H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," Automatica, June 2010, which overlaps with the present contribution in the model and criterion employed, but it is distinct in the solution, application, and real data experimentation.

Received: 31 October 2011 Accepted: 21 March 2012
Published: 21 March 2012

References

1. P Stoica, RL Moses, *Introduction to Spectral Analysis* (Prentice-Hall, New Jersey, 1997)
2. PM Djuric, A MAP solution to off-line segmentation of signals, in *Proc of the International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, **4**, 505-508 (1994)
3. N Dobigeon, J-Y Tourneret, M Davy, Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. IEEE Trans Signal Process. **55**(4), 1251–1263 (2007)
4. P Fearnhead, Exact Bayesian curve fitting and signal segmentation. IEEE Trans Signal Process. **53**(6), 2160–2166 (2005)
5. E Punskaya, C Andrieu, A Doucet, WJ Fitzgerald, Bayesian curve fitting using MCMC with applications to signal segmentation. IEEE Trans Signal Process. **50**(3), 747–758 (2002)
6. M Lavielle, Optimal segmentation of random processes. IEEE Trans Signal Process. **46**(5), 1365–1373 (1998)
7. M Basseville, IV Nikiforov, *Detection of Abrupt Changes: Theory and Application* (Prentice-Hall, Englewood Cliffs, NJ, USA, 1993)
8. MG Hall, AV Oppenheim, AS Willsky, Time-varying parametric modeling of speech. Signal Process. **5**(3), 267–285 (1983)
9. D Rudoy, TF Quatieri, PJ Wolfe, Time-varying autoregressive tests for multiscale speech analysis, in *Proceedings of Interspeech*, Brighton, UK, **5**, 2839–2842 (2009)
10. M Yuan, Y Lin, Model selection and estimation in regression with grouped variables. J Royal Stat Soc, Ser B. **68**(1), 49–67 (2006)
11. PJ Brockwell, RA Davis, *Time Series: Theory and Methods* (Springer-Verlag, New York, NY, USA, 1990)
12. L Boysen, A Kempe, V Liebscher, A Munk, O Wittich, Consistencies and rates of convergence of jump-penalized least-squares estimators. Annals Stat. **37**(1), 157–183 (2009)
13. M Lavielle, Using penalized contrasts for the change-point problem. Signal Process. **85**(8), 1501–1510 (2005)
14. M Lavielle, E Moulines, Least-squares estimation of an unknown number of shifts in a time series. J Time Series Anal. **21**(1), 33–59 (2000)
15. E Lebarbier, Detecting multiple change-points in the mean of Gaussian process by model selection. Signal Process. **85**(4), 717–736 (2005)
16. G Winkler, V Liebscher, Smoothers for discontinuous signals, J. Nonparametric Stat. **14**(1-2), 203–222 (2002)
17. Z Harchaoui, C Levy-Leduc, Catching change-points with Lasso, in *Proceedings of the Advanced Neural Information Processes Systems*, Vancouver, Canada, **20**, 161–168 (2008)
18. U Appel, AV Brandt, Adaptive sequential segmentation of piecewise stationary time series. Inf Sci. **29**(1), 27–56 (1983)
19. A Willsky, H Jones, A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. IEEE Trans Autom Control. **21**(1), 108–112 (1976)
20. YC Eldar, M Mishali, Block sparsity and sampling over a union of subspaces, in *Proc of the 16th International Conference on Digital Signal Processing*, Santorini, Greece, **1**, 1–8 (2009)
21. R Tibshirani, Regression shrinkage and selection via the Lasso. J Royal Stat Soc Ser B. **58**(1), 267–288 (1996)
22. J Fan, R Li, Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. **96**, 1348–1360 (2001)
23. B Efron, T Hastie, I Johnstone, R Tibshirani, Least angle regression. Annals Stat. **32**, 407–499 (2004)
24. JF Sturm, Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optim Methods Softw. **11-12**, 625–653 (1999)
25. A Puig, A Wiesel, A Hero, A multidimensional shrinkage-thresholding operator, in *Proceedings of the 15th Workshop on Statistical Signal Processing*, Cardiff, UK, **18**, 363–366 (2009)
26. P Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization. J Optim Theory Appl. **109**(3), 475–494 (2001)
27. YC Eldar, P Kuppinger, H Bölcskei, Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans Signal Process*. **58**, 3042–3054 (2010)
28. FR Bach, Consistency of the group Lasso and multiple kernel learning. J Mach Learn Res. **9**, 1179–1225 (2008)
29. Y Nardi, A Rinaldo, On the asymptotic properties of the group Lasso estimator for linear models. Electron J Stat. **2**, 605–633 (2008)
30. J-J Fuchs, On sparse representations in arbitrary redundant bases. IEEE Trans. Inf Theory. **50**(6), 1341–1344 (2004)
31. I Gorodnitsky, B Rao, Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. IEEE Trans Signal Process. **45**(3), 600–616 (1997)
32. J Tropp, Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans Inf Theory. **52**(3), 1030–1051 (2006)
33. EJ Candes, MB Wakin, S Boyd, Enhancing sparsity by reweighted ℓ_1 minimization. J Fourier Anal Appl. **14**(5), 877–905 (2008)
34. H Zou, R Li, One-step sparse estimates in nonconcave penalized likelihood models. Annals Stat. **36**, 1509–1533 (2008)
35. R Tibshirani, M Saunders, S Rosset, J Zhu, k Knight, Sparsity and smoothness via the fused Lasso. J Royal Stat Soc Ser B. **67**(1), 91–108 (2005)
36. A Ruszczynski, *Nonlinear Optimization* (Princeton University Press, Princeton, NJ, USA, 2006)

doi:10.1186/1687-6180-2012-70
Cite this article as: Angelosante and Giannakis: **Group lassoing change-points in piecewise-constant AR processes.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:70.