An SDP approach for ℓ_0-minimization: application to ARX model segmentation

Dario Piga [a], Roland Tóth [a]

[a] *Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Postbus 513, 5600 MB Eindhoven, The Netherlands.*

Abstract

Minimizing the ℓ_0-seminorm of a vector under convex constraints is a combinatorial (NP-hard) problem. Replacement of the ℓ_0-seminorm with the ℓ_1-norm is a commonly used approach to compute an approximate solution of the original ℓ_0-minimization problem by means of convex programming. In the theory of compressive sensing, the condition that the sensing matrix satisfies the Restricted Isometry Property (RIP) is a sufficient condition to guarantee that the solution of the ℓ_1-approximated problem is equal to the solution of the original ℓ_0-minimization problem. However, the evaluation of the conservativeness of the ℓ_1-relaxation approaches is recognized to be a difficult task in case the RIP is not satisfied. In this paper, we present an alternative approach to minimize the ℓ_0-norm of a vector under given constraints. In particular, we show that an ℓ_0-minimization problem can be relaxed into a sequence of semidefinite programming problems, whose solutions are guaranteed to converge to the optimizer (if unique) of the original combinatorial problem also in case the RIP is not satisfied. Segmentation of ARX models is then discussed in order to show, through a relevant problem in system identification, that the proposed approach outperforms the ℓ_1-based relaxation in detecting piece-wise constant parameter changes in the estimated model.

Key words: Compressive sensing, ℓ_0-minimization, Regularization, SDP relaxation, Sparse estimation, Segmentation

1 Introduction

Finding a sparse solution which satisfies a set of equality or inequality constraints is a relevant problem in many engineering areas. A typical example is the selection of a suitable model structure in linear regression, where one has to select the relevant regressors, i.e., the associated non-negligible parameters to be estimated, from a large set of possible regressors [23]. Other examples come from the *compressive sensing* (CS) theory, where a sparse signal has to be recovered from a fewer number of samples than what is required by traditional sampling methods based on Shannon's theorem (see, e.g., [4,12,7] for details on the theory of compressive sensing). Recent applications of the CS paradigm are in the field of compressive imaging [13], robotics [25], control [3], and geophysical data analysis [18].

Mathematically, the problem of finding the sparsest solution satisfying a system of linear equations $A\theta = b$ with more unknowns than equalities can be formulated as

$$\min_{\theta\in\mathbb{R}^n} \|\theta\|_0 \quad \text{s.t.} \quad A\theta = b, \tag{1}$$

where $A \in \mathbb{R}^{m,n}$, with $m < n$, $b \in \mathbb{R}^m$ and $\|\cdot\|_0$ denotes the ℓ_0-seminorm, which gives the number of nonzero components of its argument. In case the entries of the matrix A or the components of the vector b are affected by noise, the equality constraint $A\theta = b$ can be relaxed and an error tolerance ε can be allowed in the equation $A\theta = b$. This leads to the following variation of Problem (1):

$$\min_{\theta\in\mathbb{R}^n} \|\theta\|_0 \quad \text{s.t.} \quad \|A\theta - b\|_2^2 \leq \varepsilon. \tag{2}$$

Problem (2) is often considered in system identification [32,22,30,33]. In fact, consider the case when b is the vector of the (noise corrupted) output observations of a system $\mathcal{S}_o$ to be identified, A is the regression matrix, and θ is the parameter vector defining a model structure for $\mathcal{S}_o$ in the form of

$$b = A\theta + e, \tag{3}$$

⋆ This paper was not presented at any IFAC meeting. Corresponding author Dario Piga. Tel. +31-6-27-3831-3307.
⋆⋆This work was supported by the Netherlands Organization for Scientific Research (NWO, grant. no.: 639.021.127) and by the French ministries of Foreign Affairs, Education and Research and the French-Dutch Academy (PHC Van Gogh project, n. 29342QL).

Email addresses: D.Piga@tue.nl (Dario Piga), R.Toth@tue.nl (Roland Tóth).

with e being the residual vector. Then, Problem (2) aims at computing the sparsest parameter vector θ under the constraint that the ℓ_2-loss function of the prediction error in the form of (3) (i.e., $\|A\theta - b\|_2^2$) is lower than a given error threshold ε. The term ε is then tuned by the user to balance the tradeoff between sparsity of θ and minimization of $\|A\theta - b\|_2^2$. Note that, Problem (2) can be also written in the *Lagrangian form*:

$$\min_{\theta \in \mathbb{R}^n} \|A\theta - b\|_2^2 + \gamma\|\theta\|_0, \tag{4}$$

with $\gamma > 0$ as a tuning parameter. This means that for a given value of ε, there exists a value of γ such that the minima of (2) and (4) are equal.

Unfortunately, Problems (1), (2) and (4) are NP-hard and they are difficult to solve in practice for large values of n. Several methods have been developed to compute an approximate solution of such problems. More precisely, nonlinear optimization approaches which aim at computing (local) minimizers of nonconvex problems approximating (1), (2) and (4) are proposed in [9,14], while efficient greedy algorithms based on the so-called *matching pursuit* approach are discussed in [24,11,34,35]. Another common method to compute an approximate solution of (1), (2) and (4) is based on the replacement of the ℓ_0-seminorm of the vector θ with its ℓ_1-norm. This leads to the minimization of the following convex approximations of (1), (2) and (4):

$$\min_{\theta \in \mathbb{R}^n} \|\theta\|_1 \quad \text{s.t.} \quad A\theta = b, \tag{5}$$

$$\min_{\theta \in \mathbb{R}^n} \|\theta\|_1 \quad \text{s.t.} \quad \|A\theta - b\|_2^2 \leq \varepsilon, \tag{6}$$

$$\min_{\theta \in \mathbb{R}^n} \|A\theta - b\|_2^2 + \gamma\|\theta\|_1, \tag{7}$$

whose solution can be computed in polynomial time by means of convex programming techniques. Evaluation of the level of approximation introduced by Problems (5)-(7) is recognized to be a difficult task. Some results in this topic are given in [6,5], where it is shown that the solution of (5) is equal to the solution of (1) under the favorable conditions that θ is sufficiently sparse and the matrix A obeys the so-called *restricted isometry property* (RIP) [6]. However, such conditions are not satisfied in application areas such as system identification, due to the correlation of the columns of A (see, e.g., [33,31]). In this paper, we propose an alternative approach to solve problems (1), (2) and (4). First, we show how such problems can be equivalently written as polynomial optimization problems with bilinear equality constraints. Then, by using recent results proposed in [19,10,29] on convex relaxations of semialgebraic optimization problems, the solution of the formulated polynomial problems is computed by constructing a sequence of convex *semidefinite programming* (SDP) problems, whose optima are guaranteed to converge to the global optimum of (1), (2) and (4). This means that no structural approximation of the ℓ_0-norm is introduced in solving the original ℓ_0-minimization problem, and its solution can be then computed with arbitrary precision by means of convex programming. Indeed, the latter is the main advantage of the approach presented in this paper with respect to the sparse approximation methods available in literature. To summarize, the main contributions of the paper are:

- Showing that Problems (1), (2) and (4) can be approximated in terms of SDP problems with arbitrary precision.
- Illustrating the applicability of the approach on ARX segmentation.

The paper is organized as follows: the main notation used throughout the paper is defined in Section 2. In Section 3, we show how ℓ_0-minimization problems can be formulated in terms of polynomial optimization and how they can be solved by exploiting convex relaxation techniques based on the theory of moments. In Section 4, the presented method is applied to the problem of segmentation of ARX models. The obtained results are compared, by means of a simulation example, with the identification approach proposed in [27].

2 Notation

The following notation will be used throughout the paper.

$M \succeq 0$	matrix M is positive semidefinite.
$[M]_{i,j}$	(i,j)-entry of matrix M.
$\mathbb{N}_0^n$	set of n-dimensional vectors with nonnegative components.
$\mathbb{I}_m^n$	set of indices defined as $\{m, m+1, \ldots, n\}$.
x_i	i-th component of the vector $x \in \mathbb{R}^n$.
$x(\mathcal{I})$	subset of variables x_i with index i in $\mathcal{I} \subseteq \mathbb{I}_1^n$, i.e., $x(\mathcal{I}) = \{x_i \mid i \in \mathcal{I}\}$.
x^α	shorthand notation for $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} = \prod_{i=1}^n x_i^{\alpha_i}$.
$\mathcal{A}_d^n$	set defined as $\left\{\alpha \in \mathbb{N}_0^n : \sum_{i=1}^n \alpha_i \leq d\right\}$.
$\mathbb{R}_d^n[x]$	set of real-valued polynomials in the indeterminant $x \in \mathbb{R}^n$ with maximum degree d.
h_d^n	canonical basis of $\mathbb{R}_d^n[x]$, i.e., $h_d^n = [1\ x_1\ \cdots\ x_n\ x_1^2\ \cdots\ x_1 x_n\ \ x_2^2\ x_2 x_3\ \cdots\ x_n^2\ \cdots\ x_1^3\ \cdots\ x_n^d]^\top$.

3 SDP-relaxation for ℓ_0-minimization

In this section, we show that the combinatorial problem (1) can be reformulated as a constrained polynomial optimization problem whose solution can be accurately approximated by means of convex relaxation techniques based on the theory of moments. Similar considerations hold when the problem of computing a solution of (2) and (4) is considered.

3.1 Polynomial formulation of ℓ_0-minimization

Let us denote with θ^* and f^* the minimizer and the minimum of Problem (1), respectively, i.e.,

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^n} \|\theta\|_0 \quad \text{s.t.} \quad A\theta = b, \tag{8}$$

and $f^* = \|\theta^*\|_0$. An equivalent formulation of the combinatorial problem (1) is given by the following proposition.

Proposition 1 [The ℓ_0 polynomial form] Consider the optimization problem:

$$(\theta^{*,\mathrm{p}}, w^{*,\mathrm{p}}) = \arg\min_{\theta, w \in \mathbb{R}^n} \sum_{i=1}^{n} w_i \tag{9a}$$

$$\text{s.t.} \quad g_j(\theta, w) = A_j\theta - b_j = 0, \qquad j \in \mathbb{I}_1^m, \tag{9b}$$

$$g_{m+i}(\theta, w) = (1 - w_i)\theta_i = 0, \qquad i \in \mathbb{I}_1^n, \tag{9c}$$

$$g_{m+n+i}(\theta, w) = w_i \geq 0, \qquad i \in \mathbb{I}_1^n, \tag{9d}$$

with A_j denoting the j-th row of A. Problems (1) and (9) are equivalent in the following sense:

(i) they share the same minimum, i.e.,

$$f^* = f^{*,\mathrm{p}} = \sum_{i=1}^{n} w_i^{*,\mathrm{p}};$$

(ii) if θ^* is a global minimizer of (1), then

$$\theta^{*,\mathrm{p}} = \theta^* \ \text{ and } \ w_i^{*,\mathrm{p}} = \begin{cases} 1 & \text{if} \quad \theta_i^* \neq 0; \\ 0 & \text{if} \quad \theta_i^* = 0; \end{cases} \tag{10}$$

is a global minimizer of (9). On the other way around, if $(\theta^{*,\mathrm{p}}, w^{*,\mathrm{p}})$ is a global minimizer of (9), then $\theta^{*,\mathrm{p}}$ is also a global minimizer of (1).

Proof First, note that any solution $(\theta^{*,\mathrm{p}}, w^{*,\mathrm{p}})$ of (9) satisfies the following condition

$$w_i^{*,\mathrm{p}} = \begin{cases} 1 & \text{if} \quad \theta_i^{*,\mathrm{p}} \neq 0; \\ 0 & \text{if} \quad \theta_i^{*,\mathrm{p}} = 0. \end{cases} \tag{11}$$

This follows from the constraint $(1 - w_i)\theta_i = 0$ in (9) defining the feasibility set. In terms of this constraint, if $\theta_i^{*,\mathrm{p}} \neq 0$ then $w_i^{*,\mathrm{p}} = 1$. On the other hand, if $\theta_i^{*,\mathrm{p}} = 0$ then, in order to minimize the objective function of (9), $w_i^{*,\mathrm{p}} = 0$. As a consequence,

$$f^{*,\mathrm{p}} = \sum_{i=1}^{n} w_i^{*,\mathrm{p}} = \|\theta^{*,\mathrm{p}}\|_0. \tag{12}$$

Indeed, the vector $\theta^{*,\mathrm{p}}$ is a feasible solution for Problem (1) since $\theta^{*,\mathrm{p}}$ satisfies the equality constraint $A\theta^{*,\mathrm{p}} = b$ (eq. (9b)). Therefore,

$$f^{*,\mathrm{p}} = \|\theta^{*,\mathrm{p}}\|_0 \geq f^*. \tag{13}$$

Let us now consider the point (θ^*, w^*), where θ^* is a minimizer of (1) and w^* is such that

$$w_i^* = \begin{cases} 1 & \text{if} \quad \theta_i^* \neq 0; \\ 0 & \text{if} \quad \theta_i^* = 0. \end{cases} \tag{14}$$

Note that $\|\theta^*\|_0 = \sum_{i=1}^{n} w_i^* = f^*$. Since the point (θ^*, w^*) is a feasible solution of Problem (9), we have that

$$f^* = \|\theta^*\|_0 \geq f^{*,\mathrm{p}}. \tag{15}$$

By combining (13) and (15), Part (i) of the proposition follows. Part (ii) of the proposition follows straightforwardly from Part (i). In fact, if θ^* is a global minimizer of (1), then $f^{*,\mathrm{p}} = f^* = \|\theta^*\|_0$. As (θ^*, w^*) is a feasible solution of Problem (9), then $\sum_{i=1}^{n} w_i^* = \|\theta^*\|_0$. Therefore, $f^{*,\mathrm{p}} = \sum_{i=1}^{n} w_i^*$, which means that the point (θ^*, w^*) is a global minimizer of (9), as stated in Part (ii) of the proposition. Based on the same considerations, it can be proven that if $(\theta^{*,\mathrm{p}}, w^{*,\mathrm{p}})$ is a global minimizer of (9), then $\theta^{*,\mathrm{p}}$ is also a global minimizer of (1). ■

Note that (9) is a polynomial optimization problem because of the product between the variables θ_i and w_i. Recently, efficient methods have been proposed in the literature to compute approximate solutions of polynomial problems by constructing a hierarchy of SDP problems of increasing size, whose optimal values converge from below to the global minimum of the original polynomial problem. Such SDP-relaxation techniques are discussed in [19,10,29] and they are based on the theory of moments and on the dual representation of nonnegative polynomials as *sum-of-squares* (SOS). In the following, we show how the theory of moments relaxation presented in [19] can be applied to compute a solution of Problem (9).

3.2 SDP-relaxation of Problem (9)

Let $Z = [\theta^\top \ w^\top]^\top \in \mathbb{R}^{2n}$ be the collection of the optimization variables involved in Problem (9). For a given integer $\delta \geq 1$, let us rewrite the objective function of Problem (9) as

$$\sum_{i=1}^{n} w_i = \sum_{\alpha \in \mathcal{A}_{2\delta}^{2n}} F_\alpha Z^\alpha, \tag{16}$$

with $\{F_\alpha\}_{\alpha \in \mathcal{A}_{2\delta}^{2n}}$ being the expansion coefficients of $\sum_{i=1}^{n} w_i$ in the polynomial basis $\{Z^\alpha = Z_1^{\alpha_1} \cdots Z_{2n}^{\alpha_{2n}}\}_{\alpha \in \mathcal{A}_{2\delta}^{2n}}$. Consider the following optimization problem

$$f_\mu^* = \inf_{\mu \in \mathcal{P}(\mathcal{S})} \int \sum_{\alpha \in \mathcal{A}_{2\delta}^{2n}} F_\alpha Z^\alpha \mu(dZ) = \inf_{\mu \in \mathcal{P}(\mathcal{S})} \sum_{\alpha \in \mathcal{A}_{2\delta}^{2n}} F_\alpha \int Z^\alpha \mu(dZ), \tag{17}$$

where $\mathcal{S}$ is the feasibility set of Problem (9), i.e., $\mathcal{S} = \{(\theta, w) : g_j(\theta, w) = 0, j \in \mathbb{I}_1^{m+n};\ g_s(\theta, w) \geq 0, s \in \mathbb{I}_{m+n+1}^{m+2n}\}$ and $\mathcal{P}(\mathcal{S})$ is the space of finite Borel signed measures with support contained in $\mathcal{S}$. According to Proposition 2.1 in [19], Problem (9) is equivalent to Problem (17), that is

(a) $f^{*,\mathrm{p}} = f_\mu^*$;

(b) if Z^* is a global minimizer of the polynomial problem (9), then $\mu^* = \mathrm{d}(Z - Z^*)$, with $\mathrm{d}(\cdot)$ being the Dirac function on $\mathbb{R}^{2n}$.

To rewrite Problem (17) in a more convenient form, introduce $p = \{p_\alpha\}_{\alpha \in \mathcal{A}^{2n}_{2\delta}}$ as the (truncated) sequence of moments up to order 2δ of a Borel probability measure μ on $\mathbb{R}^{2n}$, i.e.,

$$p_\alpha = \int Z^\alpha \mu(dZ). \tag{18}$$

Note that p_α represents the right hand side of the summands in (17). Now we can rewrite Problem (17) as

$$f^*_\mu = \inf_p \sum_{\alpha \in \mathcal{A}^{2n}_{2\delta}} F_\alpha p_\alpha \tag{19a}$$

$$\text{s.t.} \quad p \text{ is represented by a Borel measure } \mu \in \mathcal{P}(\mathcal{S}). \tag{19b}$$

It is worth noting that (19) is an optimization problem in the variable p. Lemma 1 (below) provides necessary conditions so that p is represented by a signed Borel measure on $\mathcal{S}$. First, the following definitions are introduced. Let h^{2n}_δ be the canonical basis of $\mathbb{R}^{2n}_\delta[Z]$. Define the truncated moment matrix (of order δ) associated with the measure μ as

$$M_\delta(p) = \int h^{2n}_\delta \left(h^{2n}_\delta\right)^\top \mu(\mathrm{d}Z). \tag{20}$$

Let $\{g_{s,\beta}\}_{\beta \in \mathcal{A}^{2n}_2}$ be the sequence of the coefficients of the bilinear polynomial $g_s(Z) \in \mathbb{R}^{2n}_2[Z]$ (defining the feasible set $\mathcal{S}$) in the basis h^{2n}_2, i.e.,

$$g_s(Z) = \sum_{\beta \in \mathcal{A}^{2n}_2} g_{s,\beta} Z^\beta. \tag{21}$$

Let $p_{\alpha_{i,j}}$ be the (i,j)-entry of the truncated moment matrix (of order $\delta-1$) $M_{\delta-1}$, i.e., $p_{\alpha_{i,j}} = [M_\delta(p)]_{i,j}$. Define the localizing matrix $M_{\delta-1}(g_s p)$ of order $\delta - 1$, associated with the moments p and the polynomial $g_s(Z)$, as

$$[M_{\delta-1}(g_s p)]_{i,j} = \sum_{\beta \in \mathcal{A}^{2n}_2} g_{s,\beta} p_{\alpha_{i,j}+\beta}, \tag{22}$$

with $\alpha_{i,j} + \beta$ being the componentwise sum of the vectors $\alpha_{i,j}$ and β. The size of $M_{\delta-1}(g_s p)$ is equal to the size of the moment matrix $M_{\delta-1}$. The interested reader is referred to [19] for explanatory examples on the construction of moment and localizing matrices.

Lemma 1 [Sequence of moment conditions, [21]] If p is the sequence of moments (truncated up to order 2δ) of a Borel signed measure on $\mathcal{S}$, then: $M_\delta(p) \succeq 0$; $M_{\delta-1}(g_s p) = 0$ for all $s \in \mathbb{I}^{m+n}_1$ and $M_{\delta-1}(g_s p) \succeq 0$ for all $s \in \mathbb{I}^{m+2n}_{m+n+1}$. ∎

On the basis of Lemma 1, Problem (19) can be relaxed by replacing the constraint (19b) with the less restrictive matrix constraints given in Lemma 1. This leads to the following relaxation of (19) (or equivalently of (9)):

$$\min_p \sum_{\alpha \in \mathcal{A}^{2n}_{2\delta}} F_\alpha p_\alpha \tag{23a}$$

$$\text{s.t.} \quad M_\delta(p) \succeq 0, \quad M_{\delta-1}(g_s p) = 0, \quad s \in \mathbb{I}^{m+n}_1, \tag{23b}$$

$$M_{\delta-1}(g_s p) \succeq 0, \quad s \in \mathbb{I}^{m+2n}_{m+n+1}. \tag{23c}$$

It is worth remarking that the matrices $M_\delta(p)$ and $M_{\delta-1}(g_s p)$ depend linearly on the optimization variables $p = \{p_\alpha\}_{\alpha \in \mathcal{A}^{2n}_{2\delta}}$. Therefore, Problem (23) is a convex SDP problem with linear objective function and *linear matrix inequality* (LMI) constraints in the variables p. **Example 1** For the sake of clarity, the above introduced theory of moment relaxation is explained here through a simple example. Consider the following ℓ_0-minimization problem:

$$\min_{\theta_1,\theta_2} \; \| [\theta_1 \; \theta_2]^\top \|_0 \qquad \text{s.t.} \quad a_1\theta_1 + a_2\theta_2 = b, \tag{24}$$

where a_1, a_2 and b are known constants. The polynomial problem associated with (24) is then

$$\min_{\substack{\theta_1,\theta_2\\ w_1,w_2}} \quad w_1 + w_2 \tag{25}$$

$$\begin{aligned} \text{s.t.} \;\; & g_1(\theta,w) = a_1\theta_1 + a_2\theta_2 - b = 0, \\ & g_2(\theta,w) = (1-w_1)\theta_1 = 0, \quad g_3(\theta,w) = (1-w_2)\theta_2 = 0, \\ & g_4(\theta,w) = w_1 \geq 0, \qquad\qquad g_5(\theta,w) = w_2 \geq 0. \end{aligned}$$

Based on the previously described theoretical results concerning the relaxation of (9) with (23), it follows that the SDP-relaxed problem of order $\delta = 1$ associated with (25) is given by:

$$\min_p \quad p_{0010} + p_{0001} \tag{26a}$$

$$\text{s.t.} \quad M_1(p) \succeq 0; \; M_0(g_s p) = 0, \; s \in \mathbb{I}^3_1; \tag{26b}$$

$$M_0(g_s p) \succeq 0, \; s \in \mathbb{I}^5_4. \tag{26c}$$

In (26b), the moment matrix $M_1(p)$ is

$$M_1(p) = \begin{bmatrix} p_{0000} & p_{1000} & p_{0100} & p_{0010} & p_{0001} \\ p_{1000} & p_{2000} & p_{1100} & p_{1010} & p_{1001} \\ p_{0100} & p_{1100} & p_{0200} & p_{0110} & p_{0101} \\ p_{0010} & p_{1010} & p_{0110} & p_{0020} & p_{0011} \\ p_{0001} & p_{1001} & p_{0101} & p_{0011} & p_{0002} \end{bmatrix},$$

and the localizing matrices are

$$\begin{aligned} & M_0(g_1 p) = a_1 p_{1000} + a_2 p_{0100} - b p_{0000}, \\ & M_0(g_2 p) = 1 - p_{1010}, \quad M_0(g_3 p) = 1 - p_{0101}, \\ & M_0(g_4 p) = p_{0010}, \qquad M_0(g_5 p) = p_{0001}. \end{aligned}$$ ∎

Property 1 [Convergence properties] Denote with $\tilde{f}^*$ and p^* the minimum and the minimizer, respectively, of the SDP problem (23). Let $\hat{p}^*$ be defined as

$$\hat{p}^* := \left\{ p^*_\alpha \,|\, \sum_{i=1}^{2n} \alpha_i = 1 \right\} = \{p^*_{100\cdots00}, p^*_{010\cdots00}, \ldots, p^*_{000\cdots01}\}.$$

The following asymptotic convergence results hold as δ goes to infinity:

(i) $\tilde{f}^*$ monotonically converges, from below, to the global optimum of the combinatorial Problem (1).
(ii) If the global minimizer $(\theta^{*,\mathrm{p}},\ w^{*,\mathrm{p}})$ of Problem (9) is unique, then $\hat{p}^*$ converges to $(\theta^{*,\mathrm{p}},\ w^{*,\mathrm{p}})$. ■

The features highlighted in Property 1 follow from the application of Theorem 4.2 in [19] to the polynomial optimization problem (9) and to the corresponding SDP-relaxed problem (23).

The convergence properties highlighted in Property 1 can be interpreted as follows: the global optimum of a (nonconvex) polynomial optimization problem can be computed by solving a (convex) SDP problem with infinite-size LMI constrains. To put it another way, the nonconvexity of the problem can be equivalently replaced by infinite-dimensional convex constraints. It is also important to point out that the convergence properties in Property 1 can be also achieved for finite δ. However, the minimum value of δ which guarantees convergence is not known a-priori. As a consequence, computing the exact solution of the original ℓ_0-minimization problem through the SDP problem (23) might require more time than using a brute-force algorithm. Nevertheless, the global optimum and the global minimizer of the original combinatorial problem (1) are usually attained in practice at a small value of δ (see, e.g., [16] for a collection of polynomial optimization problems solved with relaxation orders smaller than 4). Furthermore, according to [17], the numerical test

$$\mathrm{rank}(M_\delta(p^*)) = \mathrm{rank}(M_{\delta-1}(p^*)), \tag{27}$$

can be carried out in order to check if the global optimum of (9) is attained by the corresponding SDP-problem (23) at a given finite value of δ. In case condition (27) is satisfied, the global minimizer θ^* of the combinatorial Problem (1) can be extracted from $M_\delta(p^*)$ through the procedure presented in [17]. It is worth pointing out that the convergence properties stated in Property 1 (ii) are guaranteed only if the global minimizer of (9) is unique. Nevertheless, when the rank condition in (27) is satisfied, the extraction procedure in [17] provides a global minimizer of Problem (9) also in case of multiple minimizers.

Example 2 To demonstrate how the proposed approach can significantly outperform the standard ℓ_1-relaxation methods, a simple but illustrative example is provided. Consider the optimization Problem (1) with

$$A = \begin{bmatrix} 2 & -1 & 31 & 3 \\ 10/3 & 3.33 & 44.17 & 2.5 \\ -4/3 & 4.67 & -26.67 & -4 \end{bmatrix}; \qquad b = \begin{bmatrix} 6 \\ 10 \\ -4 \end{bmatrix}.$$

The minimizer $\hat{p}^*$ obtained by solving the SDP-problem (23) for a relaxation order $\delta = 2$ is equal to $\hat{p}^* = [3\ 0\ 0\ 0]^\top$, which is indeed the global minimizer θ^* of the original nonconvex problem (1). On the other hand, the optimizer $\tilde{\theta}$ of the relaxed problem (5), obtained by replacing the ℓ_0-seminorm of θ with its ℓ_1-norm, is equal to $\tilde{\theta} = \left[0\ \frac{22}{99}\ \frac{22}{99}\ -\frac{22}{99}\right]^\top$. Indeed, the computed value of $\tilde{\theta}$ is completely different from the true minimizer of Problem (1). Furthermore, the support of $\tilde{\theta}$ is the complement of the support of θ^*. Of course, one can argue that the entries of the matrix A do not allow a good approximation of Problem (1) with its corresponding ℓ_1-relaxed problem (5). Nevertheless, it is important to point out that the ℓ_1-relaxation can have a significant price in terms of the accuracy of the approximate solution. On the other hand, the solution computed through the proposed SDP-approach asymptotically converges to the exact one. Furthermore, convergence to the exact solution for a finite degree δ is usually achieved in practice and it can be verified via the numerical test (27). ■

Unfortunately, because of the high computational burden, SDP problems arising directly from the relaxation of polynomial problems like (9) can be solved in commercial workstations by means of state-of-the-art general-purpose SDP solvers, like *SeDuMi*, only if the dimension n of the vector θ is approximately not greater than 15. In order to overcome such a limitation, two different approaches can be exploited:

- Use of numerical algorithms tailored to solve structured SDP problems arising from the relaxation of polynomial problems via the theory of moments. The algorithms proposed in [26] and [36] provide promising approaches in this direction.
- Exploit structure, if present, in the original ℓ_0-minimization problem. This allows to reduce the size of the SDP problems arising from the moment-relaxation theory.

The latter approach is illustrated in the next section via an applied problem of segmentation of ARX models.

4 Segmentation of ARX models

The aim of this section is to show how to exploit the structured sparsity in ℓ_0-minimization problems in order to reduce the computational complexity of the proposed SDP-relaxation approach. In particular, inspired by [27], we show how to apply the SDP-approach to the problem of segmentation of ARX models, a well-known sparse estimation problem in system identification.

4.1 Motivation and problem description

Consider the problem of identifying a time-varying discrete-time *auto regressive with exogenous input*

(ARX) model in the form of

$$y(k) = \phi^\top(k)\theta(k) + e_\theta(k), \tag{28}$$

where $y(k) \in \mathbb{R}$ is the measured output of the data-generating system $\mathcal{S}_\mathrm{o}$ to be modelled, $\phi(k) \in \mathbb{R}^n$ is the regressor vector at time k with shifted input and output signals of $\mathcal{S}_\mathrm{o}$, $e_\theta(k)$ is the prediction error and $\theta(k) \in \mathbb{R}^n$ are the time-varying model parameters to be identified based on a measured data record $\mathcal{D}_N = \{\phi(k), y(k)\}_{k=1}^N$. It is assumed that $\mathcal{S}_\mathrm{o}$ can be represented by (28), i.e., there exists a set of parameters $\theta_\mathrm{o}(k)$ such that $e_\theta(k)$ is white. The parameters $\theta_\mathrm{o}(k)$ and hence $\theta(k)$ are assumed to be piecewise constant and not to change frequently. The time instants when the model parameters change are *a-priori* unknown. The considered identification problem, referred to as model or signal segmentation, is considered in various papers (see, e.g., [1,27,2,28]) and it has been successfully applied in real-world problems such as image representation [8] and econometric analysis of stock markets [15].

A *standard least-square* (LS) estimate can be performed by minimizing the ℓ_2-loss function $\mathcal{V}(\theta, \mathcal{D}_N)$ of the prediction in terms of (28), i.e.,

$$\mathcal{V}(\theta, \mathcal{D}_N) = \sum_{k=1}^{N} (y(k) - \phi^\top(k)\theta(k))^2. \tag{29}$$

The estimate of the model parameters is then given by the argument of:

$$\min_\theta \sum_{k=1}^{N} (y(k) - \phi^\top(k)\theta(k))^2, \tag{30}$$

where $\theta \in \mathbb{R}^{n_\theta} : \theta = [\theta^\top(1)\ \theta^\top(2)\ \cdots\ \theta^\top(N)]^\top$ and $n_\theta = nN$. However, in the time-varying case, the LS estimate leads to an overparameterized problem with nN parameters and N measurements. One possible solution to overcome such a problem is to introduce a regularization term in (30) to penalize parameter variations. Then, an estimate of the parameters θ can be obtained by solving the minimization problem

$$\min_\theta \sum_{k=1}^{N} \big(y(k) - \phi^\top(k)\theta(k)\big)^2 + \gamma\|\Delta\theta\|_0, \tag{31}$$

where $\gamma > 0$ is the so-called regularization parameter, and the k-th component $\Delta\theta_k$ of the vector $\Delta\theta \in \mathbb{R}^{N-1}$ is

$$\Delta\theta_k = \|\theta(k+1) - \theta(k)\|_q, \tag{32}$$

for $q = 1, 2$ or ∞. Note that, because of the definition of $\Delta\theta_k$ in (32), the regularization term $\|\Delta\theta\|_0$ counts the number of times when $\theta(k+1) - \theta(k)$ is different from the 0-vector (i.e., there is a variation from time k to time $k+1$ of at least one component of the parameter vector $\theta(k)$). In [27], the combinatorial problem (31) is relaxed by replacing the ℓ_0-seminorm of $\Delta\theta$ with its ℓ_1-norm. An estimate of the parameter sequence $\theta(k)$ is then computed by minimizing the convex problem

$$\min_\theta \sum_{k=1}^{N} \big(y(k) - \phi^\top(k)\theta(k)\big)^2 + \gamma\|\Delta\theta\|_1. \tag{33}$$

In this work we show how to solve Problem (31) through the approach discussed in Section 3. In particular, we show how to exploit the inherent structured sparsity of Problem (31) to reduce the size of the SDP problems arising from the theory-of-moment relaxation.

4.2 Segmentation as a sparse polynomial problem

Based on results similar to the ones reported in Proposition 1, the combinatorial problem (31) can be equivalently written as

$$\begin{aligned} &\min_{\theta, w} \sum_{k=1}^{N} \big(y(k) - \phi^\top(k)\theta(k)\big)^2 + \gamma \sum_{k=1}^{N-1} w_k \\ &\text{s.t. } g_{(k-1)n+i} = (1 - w_k)(\theta_i(k+1) - \theta_i(k)) = 0, \\ &\qquad i \in \mathbb{I}_1^n,\ k \in \mathbb{I}_1^{N-1}, \\ &\qquad g_{(N-1)n+k} = w_k \geq 0,\ k \in \mathbb{I}_1^{N-1}. \end{aligned} \tag{34}$$

Property 2 [Polynomial structure of ARX segmentation]
Problem (34) enjoys the following features:

(i) The term $\big(y(k) - \phi^\top(k)\theta(k)\big)^2$ appearing in the objective function depends only on the parameters $\theta(k)$.
(ii) For every $k \in \mathbb{I}_1^{N-1}$ and for every $i \in \mathbb{I}_1^n$, the constraints $g_{(k-1)n+i} = 0$ are bilinear and only depend on the variable w_k and on the parameters $\theta_i(k)$ and $\theta_i(k+1)$.
(iii) For every $k \in \mathbb{I}_1^{N-1}$, the constraint $g_{(N-1)n+k} \geq 0$ depends only on the variable w_k. ■

Thanks to Property 2, a structural pattern in Problem (34) can be easily detected and thus used to formulate a reduced version of the theory of moments relaxation as described in the following. Collect the optimization variables of Problem (34) into the vector $X = \big[\theta^\top(1)\ \ldots\ \theta^\top(N)\ w^\top\big]^\top \in \mathbb{R}^{nN+N-1}$. For $k \in \mathbb{I}_1^{N-1}$, define the index sets $\mathcal{I}_k \subset \mathcal{I}_0 = \mathbb{I}_1^{nN+N-1}$ and $\mathcal{S}_k \subset \mathcal{S}_0 = \mathbb{I}_1^{(n+1)(N-1)}$ as

$$\mathcal{I}_k = \{(k-1)n + i, kn + i, Nn + k,\ i \in \mathbb{I}_1^n\}, \tag{35a}$$

$$\mathcal{S}_k = \{(k-1)n + 1,\ \ldots,\ (k-1)n + n,\ (N-1)n + k\}. \tag{35b}$$

The index sets $\mathcal{I}_k$ and $\mathcal{S}_k$ are constructed on the basis of the structure in Problem (34) highlighted by Property 2. More precisely, the sets $\mathcal{I}_k$ and $\mathcal{S}_k$ are defined so that, for all $s \in \mathcal{S}_k$, the polynomial g_s in (34) only depend on $\theta_i(k), \theta_i(k+1)$ (with $i \in \mathbb{I}_1^n$) and w_k, which are precisely the variables X_i with $i \in \mathcal{I}_k$. The following property highlights the features of the index sets $\mathcal{I}_k$ and $\mathcal{S}_k$ which will play a crucial role in guaranteeing the key results given in Proposition 4.

Property 3 [Features of the index sets]
For every $k \in \mathbb{I}_1^{N-1}$, the index sets $\mathcal{I}_k$ and $\mathcal{S}_k$ are such that:

(i) $\mathcal{I}_0 = \bigcup_{k=1}^{N-1} \mathcal{I}_k$.

(ii) $\mathcal{S}_0 = \bigcup_{k=1}^{N-1} \mathcal{S}_k$.

(iii) The sets $\mathcal{S}_k$ are mutually disjoint, i.e., $S_k \cap S_j = \emptyset$ if $k \neq j$.

(iv) For every $s \in \mathcal{S}_k$, the generic polynomial g_s defining the feasibility set of Problem (34) only depends on the variables $X(\mathcal{I}_k) = \{X_i : i \in \mathcal{I}_k\}$.

(v) For every $k \in \mathbb{I}_1^{N-2}$, the set $\mathcal{I}_{k+1}$ is such that

$$\mathcal{I}_{k+1} \cap \bigcup_{j=1}^{k} \mathcal{I}_j \subseteq \mathcal{I}_k.$$

(vi) The objective function of Problem (34) can be written as

$$\sum_{k=1}^{N-1} f_k\left(X(\mathcal{I}_k)\right), \tag{36}$$

where $f_k\left(X(\mathcal{I}_k)\right)$ is a polynomial function involving only the variables $X(\mathcal{I}_k)$. More precisely:

$$\begin{aligned}
f_k\big(X(\mathcal{I}_k)\big) &= \big(y(k) - \phi^\top(k)\theta(k)\big)^2 + \gamma w_k \;\text{ if }\; k \in \mathbb{I}_1^{N-2},\\
f_k\left(X(\mathcal{I}_k)\right) &= \big(y(N-1) - \phi^\top(N-1)\theta(N-1)\big)^2 + \\
&\quad + \big(y(N) - \phi^\top(N)\theta(N)\big)^2 + \gamma w_{N-1} \text{ if } k = N-1
\end{aligned}$$

■

For a given relaxation order $\delta \geq 1$, let $p = \{p_\alpha\}_{\alpha \in \mathcal{A}_{2\delta}^{nN+N-1}}$ be the sequence of moments up to order 2δ associated with the decision variables X of Problem (34). Based on the ideas presented in [20], a sparse version of the SDP-relaxation approach discussed in Section 3 can be applied to relax (34) into the following convex SDP problem

$$\begin{aligned}
&\min_{p} \sum_{\alpha \in \mathcal{A}_{2\delta}^{nN+N-1}} C_\alpha p_\alpha && (37)\\
&\text{s.t. } M_\delta(p, \mathcal{I}_k) \succeq 0, \qquad k \in \mathbb{I}_1^{N-1},\\
&\quad M_{\delta-1}(g_s p, \mathcal{I}_k) = 0,\; s \in \Big(\mathcal{S}_k \cap \mathbb{I}_1^{(N-1)n}\Big),\; k \in \mathbb{I}_1^{N-1},\\
&\quad M_{\delta-1}(g_s p, \mathcal{I}_k) \succeq 0,\; s \in \Big(\mathcal{S}_k \cap \mathbb{I}_{(N-1)n+1}^{(N-1)(n+1)}\Big),\; k \in \mathbb{I}_1^{N-1},
\end{aligned}$$

where $C = \{C_\alpha\}_{\alpha \in \mathcal{A}_{2\delta}^{nN+N-1}}$ is the coefficient vector of the objective function in (34) in the basis $\{X^\alpha\}_{\alpha \in \mathcal{A}_{2\delta}^{nN+N-1}}$, $M_\delta(p, \mathcal{I}_k)$ is the moment matrix of order δ associated with the variables $X(\mathcal{I}_k)$ and $M_{\delta-1}(g_s p, \mathcal{I}_k)$ is the localizing matrix associated with the variables $X(\mathcal{I}_k)$ realizing the constraint $g_s \geq 0$ which defines the feasibility set of Problem (34). The moment matrices $M_\delta(p, \mathcal{I}_k)$ (resp. the localizing matrix $M_{\delta-1}(g_s p, \mathcal{I}_k)$) can be obtained by retaining only those rows and columns of the moment matrix $M_\delta(p)$ (resp. of the localizing matrix $M_{\delta-1}(g_s p)$) where the variables p_α are such that $\mathrm{Supp}(\alpha) \subseteq \mathcal{I}_r$, with $\mathrm{Supp}(\alpha)$ denoting the support of α. The interested reader is referred to the papers [20,37] for more details on the construction of moment and localizing matrices for structured polynomial problem relaxation.

Results similar to the ones stated in Property 1 hold, according to the following proposition.

Property 4 [Convergence of the SDP problems] Let $\tilde{f}^*$ and p^* be the minimum and the minimizer, respectively, of the SDP problem (37). Let $\hat{p}^*$ be defined as $\hat{p}^* := \{p_\alpha^* \mid \sum_{i=1}^{nN+N-1} \alpha_i = 1\}$, i.e., $\hat{p}^* := \{p^*_{100\cdots00}, p^*_{010\cdots00}, \ldots, p^*_{000\cdots01}\}$. The following asymptotic convergence results hold as δ goes to infinity:

(i) $\tilde{f}^*$ monotonically converges, from below, to the global optimum of the original nonconvex problem (31);

(ii) $\hat{p}^*$ converges to the global minimizer (if unique) of Problem (34).

Furthermore, if for a finite value of δ

$$\mathrm{rank}(M_\delta(p^*, \mathcal{I}_k)) = \mathrm{rank}(M_{\delta-1}(p^*, \mathcal{I}_k)),\; \forall k \in \mathbb{I}_1^{N-1}, \tag{38}$$

and $\mathrm{rank}(M_{\delta-1}(p^*, \mathcal{I}_{k,q})) = 1$ for all pairs (k,q) such that $\mathcal{I}_{k,q} = \mathcal{I}_k \cap \mathcal{I}_q \neq \emptyset$, then $\tilde{f}^*$ is equal to the global minimum of Problem (34). ■

Statement of the Property follows from the application of Theorems 3.6 and 3.7 in [20] to (34) and to the corresponding SDP-relaxed problem (37). Note that the features enjoyed by the sets $\mathcal{I}_k$ and $\mathcal{S}_k$ and highlighted by Property 3 are crucial to apply the theorems in [20] to problems (34) and (37).

It is important to remark that, because of computational complexity, the developed SDP-approach can only be applied for relaxation orders $\delta \leq 3$. Nevertheless, since only linear and bilinear constraints are involved in (34), a good approximation of the exact solution of (34) is obtained in practice (based on the authors' experience) for low relaxation orders, i.e., $\delta = 2$.

4.3 A simulation example

In the sequel, a simulation example is given in order to show the effectiveness of the presented approach in the segmentation of ARX models. A comparison with the method proposed in [27] is also provided.

The behavior of the data-generating system $\mathcal{S}_{\mathrm{o}}$ is described by the difference equation

$$\begin{aligned}
y(k) = &\theta_1^{\mathrm{o}}(k)y(k-1) + \theta_2^{\mathrm{o}}(k)y(k-2) + \theta_3^{\mathrm{o}}(k)u(k) + \\
&+ \theta_4^{\mathrm{o}}(k)u(k-1) + \theta_5^{\mathrm{o}}(k)u(k-2) + e_{\mathrm{o}}(k),
\end{aligned} \tag{39}$$

with parameters variations

$$\theta_1^{\mathrm{o}} = 0.6,\; \theta_2^{\mathrm{o}} = 0.3,\; \theta_3^{\mathrm{o}} = 1,\; \theta_4^{\mathrm{o}} = 0.5,\; \theta_5^{\mathrm{o}} = -0.3,$$

from $k = 1$ to $k = 60$;

$$\theta_1^{\mathrm{o}} = 0.7,\; \theta_2^{\mathrm{o}} = -0.6,\; \theta_3^{\mathrm{o}} = 1.4,\; \theta_4^{\mathrm{o}} = -0.4,\; \theta_5^{\mathrm{o}} = -0.2,$$

from $k = 61$ to $k = 130$,

$$\theta_1^{\mathrm{o}} = -0.5,\; \theta_2^{\mathrm{o}} = -0.4,\; \theta_3^{\mathrm{o}} = -0.2,\; \theta_4^{\mathrm{o}} = 1.3,\; \theta_5^{\mathrm{o}} = 0.9,$$

from $k = 131$ to $k = 200$.

The input $u(k)$ is a (zero-mean) white noise sequence with uniform distribution $\mathcal{U}(-1, 1)$ and $e_o(k)$ is a white noise process with normal distribution $\mathcal{N}(0, \sigma_e^2)$ and standard deviation $\sigma_e = 0.05$. This corresponds to a *signal to noise ratio* (SNR) of 24 dB. The chosen model structure is defined in the form of (39), i.e.,

$$\begin{aligned} y(k) =& \theta_1^o(k)y(k-1) + \theta_2(k)y(k-2) + \theta_3(k)u(k) + \\ &+ \theta_4(k)u(k-1) + \theta_5(k)u(k-2) + e(k), \end{aligned} \tag{40}$$

The parameter sequence $\theta(k)$ is first estimated through the ℓ_1-relaxation approach proposed in [27] by approximating (31) with the convex problem (33). In order to improve the estimate, the iterative refinement algorithm discussed in [27] is applied. The regularization parameter γ is set to 50. The value of γ has been selected using an exhaustive grid search in order to have 3 segments in the estimates of the parameters. The time instants when a model change occurs are detected and, by exploiting this information, a final least-square fit is carried out by constraining the system parameters to be constant over fixed time intervals. Next, the time instants when the parameter changes occur are estimated by using the proposed SDP approach, solving Problem (37) for a relaxation order $\delta = 2$ and for $\gamma = 50$. By means of the global optimality check described in Proposition 4, the values of the parameters computed by solving (37) are proven to be the global minimizer of the original combinatorial problem (31). An extra least-squares fit is then performed to estimate the values of the ARX model parameters. The CPU time taken by the Matlab SDP-solver SeDuMi to compute the solution of Problem (37) is 2014 seconds on a 2.40-GHz Intel Pentium IV with 3 GB of RAM, while the time taken to compute the parameter estimates through the ℓ_1 based method is 9 seconds. The ℓ_1-norm $\|\hat{\theta}_i - \theta_i^o\|_1$ (with $i = 1, \ldots, 5$) of the difference between the true parameters θ_i^o and the estimated parameters $\hat{\theta}_i$ is reported in Table 1. The estimates of the parameters $\hat{\theta}_1$ and $\hat{\theta}_5$ are plotted in Fig. 1 and Fig. 2. The ℓ_1-relaxation technique provides an estimate of the parameter variations at time $k = 128$, while the true parameters vary at time $k = 131$. On the other hand, the time instants when parameter variations occur are correctly detected by the SDP-relaxation approach. We want to point out that, by our experience, the identification technique proposed in [27] provides, in many examples, an exact estimate of the time instants when model changes occur. Therefore, the SDP approach discussed in the paper should not necessarily be seen as an improvement of [27], but as a promising alternative algorithm for segmentation of ARX models in cases when the ℓ_1 relaxation introduces a significant error.

5 Conclusions

A novel approach to handle the combinatorial problem of minimizing the ℓ_0-seminorm of a vector over a set of constraints is presented in this paper. First, ℓ_0-seminorm based minimization is formulated in terms of polynomial optimization. Then, by using recent results on the optimization of semialgebraic problems through the theory of moments relaxation, the optimal solution of the formulated polynomial problem is computed by constructing a sequence of relaxed semidefinite programming (SDP) problems, whose optima converge to the global optimum of the original ℓ_0-seminorm minimization problem. An example is presented in the paper to show that, when the restricted isometry property is not satisfied, ℓ_1-relaxation based methods can fail in computing the sparsest solution of an underdetermined system of linear equations. On the other hand, the proposed SDP approach is able to provide the exact solution (if unique) of the considered problem. The presented method is applied to the segmentation of ARX models, where the inherent structured sparsity of the considered identification problem is exploited to reduce the computational complexity of the proposed approach. The main advantage of the SDP-relaxation method with respect to the other approximation approaches available in the literature is that the optimal solution of the original ℓ_0-minimization problem can be computed with arbitrary precision by means of convex programming techniques. Furthermore, a numerical certificate can be checked to detect if the global optimum of the ℓ_0-minimization problem is attained by the corresponding SDP-relaxed problem at a given relaxation order.

References

[1] P. Andersson. Adaptive forgetting in recursive identification through multiple models. *Int. Journal of Control*, 42(5):1175–1193, 1985.

[2] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Trans. on Automatic Control*, 50(10):1567–1580, 2005.

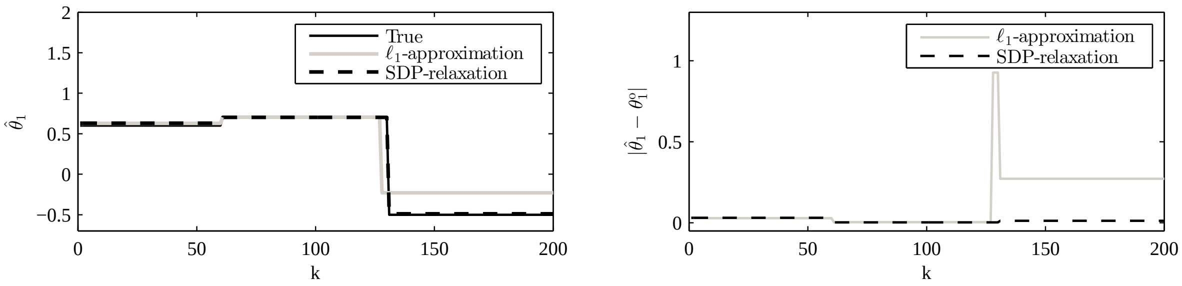
[3] S. Bhattacharya and T. Basar. Sparsity based feedback design: A new paradigm in opportunistic sensing. In *Proc. of the American Control Conference*, pages 3704–3709, San Francisco, California (USA), 2011.

[4] E.J. Candès. Compressive sampling. In *Proc. of the International Congress of Mathematicians*, pages 1433–1452, Madrid, Spain, 2006.

[5] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52(2):489–509, 2006.

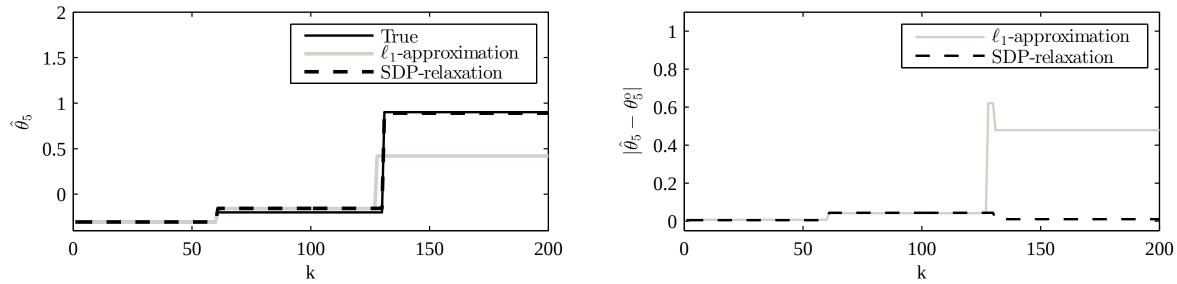
Table 1
ℓ_1-norm of the parameter estimation error.

	ℓ_1-approximation	SDP-relaxation
$\|\hat{\theta}_1 - \theta_1^o\|_1$	23.7	2.8
$\|\hat{\theta}_2 - \theta_2^o\|_1$	5.1	3.2
$\|\hat{\theta}_3 - \theta_3^o\|_1$	16.1	0.5
$\|\hat{\theta}_4 - \theta_4^o\|_1$	9.6	2.9
$\|\hat{\theta}_5 - \theta_5^o\|_1$	38.6	4.1



(a) (b)

Fig. 1. (a) Estimates of the parameter sequence $\theta_1(k)$; (b) Difference (in absolute value) between the true parameter sequence $\theta_1^o(k)$ and the estimated parameter sequence $\hat{\theta}_1(k)$.



(a) (b)

Fig. 2. (a) Estimates of the parameter sequence $\theta_5(k)$; (b) Difference (in absolute value) between the true parameter sequence $\theta_5^o(k)$ and the estimated parameter sequence $\hat{\theta}_5(k)$.

[6] E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51(12):4203–4215, 2005.

[7] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[8] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[9] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[10] G. Chesi, A. Garulli, A. Tesi, and A. Vicino. Solving quadratic distance problems: an LMI-based approach. *IEEE Trans. on Automatic Control*, 48(2):200–212, 2003.

[11] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[12] D.L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[13] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K.F. Kelly, and R.G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.

[14] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. on Signal Processing*, 57(12):4686–4698, 2009.

[15] B. Guo, Y. Wu, H. Xie, and B. Miao. A segmented regime-switching model with its application to stock market indices. *Journal of Applied Statistics*, 38(10):2241–2252, 2011.

[16] D. Henrion and J.B. Lasserre. Solving nonconvex optimization problems. *IEEE Control Systems Magazine*, 24(3):72–83, 2004.

[17] D. Henrion and J.B. Lasserre. Detecting global optimality and extracting solutions in gloptipoly. In *Positive Polynomials in Control*, volume 312 of *Lecture Notes in Control and Information Sciences*, pages 581–581. Springer Heidelberg, 2005.

[18] F.J. Herrmann and G. Hennenfent. Non-parametric seismic data recovery with curvelet frames. *Geophysical Journal International*, 173(1):233–248, 2008.

[19] J.B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11:796–817, 2001.

[20] J.B. Lasserre. Convergent semidefinite relaxations in polynomial optimization with sparsity. *SIAM Journal on Optimization*, 17(1):822–843, 2006.

[21] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. *Emerging Applications of Algebraic Geometry, Vol. 149*, pages 157–270, 2009.

[22] I. Lind and L. Ljung. Regressor and structure selection in NARX models using a structured ANOVA approach. *Automatica*, 44(2):383–395, 2008.

[23] L. Ljung. *System Identification, Theory for the User.* Prentince Hall, Upper Saddle River, 1999.

[24] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415, 1993.

[25] M. Nagahara and D.E. Quevedo. Sparse representations for packetized predictive networked control. In *Proc. of the 18th IFAC World Congress*, pages 84–89, Milan, Italy, 2011.

[26] J. Nie and L. Wang. Regularization methods for SDP relaxations in large-scale polynomial optimization. *SIAM*

Journal on Optimization, 22(2):408–428, 2012.

[27] H. Ohlsson, L. Ljung, and S. Boyd. Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6):1107–1111, 2010.

[28] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of switched affine systems. *IEEE Transactions on Automatic Control*, 99(3):634–648, 2012.

[29] P. Parrillo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96:293–320, 2003.

[30] C.R. Rojas and H. Hjalmarsson. Sparse estimation based on a validation criterion. In *Proc. of the 50th IEEE Conference on Decision and Control*, pages 2825–2830, Orlando, Florida (USA), 2011.

[31] B.M. Sanandaji, T. L. Vincent, and M. B. Wakin. Concentration of measure inequalities for compressive toeplitz matrices with applications to detection and system identification. In *Proc. of the 49th IEEE Conference on Decision and Control*, pages 2922–2929, Atlanta, Georgia (USA), 2010.

[32] R. Tibshirani. Regression shrinkage and selection via the lasso. *Jornal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[33] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent. Compressive system identification in the linear time-invariant framework. In *Proc. of the 50th IEEE Conference on Decision and Control*, pages 783–790, Orlando, Florida (USA), 2011.

[34] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. on Information Theory*, 50(10):2231–2242, 2004.

[35] J.A. Tropp, A.C. Gilbert, and M.J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

[36] C. Vo, M. Muramatsu, and M. Kojima. Equality based contraction of semidefinite programming relaxations in polynomial optimization. *Journal of the Operations Research Society of Japan*, 51(1):111–125, 2008.

[37] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. Sums of squares and semidefinite programming relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17(1):218–242, 2006.