2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

# Compressive System Identification of LTI and LTV ARX Models

Borhan M. Sanandaji,⋆ Tyrone L. Vincent,⋆ Michael B. Wakin,⋆ Roland Tóth,† and Kameshwar Poolla⋄

*Abstract*— **In this paper, we consider identifying Auto Regressive with eXternal input (ARX) models for both Linear Time-Invariant (LTI) and Linear Time-Variant (LTV) systems. We aim at doing the identification from the *smallest possible* number of observations. This is inspired by the field of Compressive Sensing (CS), and for this reason, we call this problem Compressive System Identification (CSI).**

**In the case of LTI ARX systems, a system with a large number of inputs and unknown input delays on each channel can require a model structure with a large number of parameters, unless input delay estimation is performed. Since the complexity of input delay estimation increases exponentially in the number of inputs, this can be difficult for high dimensional systems. We show that in cases where the LTI system has possibly many inputs with different unknown delays, simultaneous ARX identification and input delay estimation is possible from few observations, even though this leaves an apparently ill-conditioned identification problem. We discuss identification guarantees and support our proposed method with simulations.**

**We also consider identifying LTV ARX models. In particular, we consider systems with parameters that change only at a few time instants in a piecewise-constant manner where neither the change moments nor the number of changes is known a priori. The main technical novelty of our approach is in casting the identification problem as recovery of a block-sparse signal from an underdetermined set of linear equations. We suggest a random sampling approach for LTV identification, address the issue of identifiability and again support our approach with illustrative simulations.**

## I. Introduction

Classical system identification approaches have limited performance in cases when the number of available data samples is small compared to the order of the system [1]. These approaches usually require a large data set in order to achieve a certain performance due to the asymptotic nature of their analysis. On the other hand, there are many application fields where only limited data sets are available. Online estimation, Linear Time-Variant (LTV) system identification and setpoint-operated processes are examples of situations for which limited data samples are available. For some specific applications, the cost of the measuring process or the computational effort is also an issue. In such situations, it is *necessary* to perform the system identification from the smallest possible number of observations, although doing so leaves an apparently ill-conditioned identification problem. However, many systems of practical interest are less complicated than suggested by the number of parameters in a standard model structure. They are often either low-order or can be represented in a suitable basis or formulation in which the number of parameters is small. The key element is that while the proper representation may be known, the particular elements within this representation that have non-zero coefficients are unknown. Thus, a particular system may be expressed by a coefficient vector with only a few non-zero elements, but this coefficient vector must be high-dimensional because we are not sure a priori which elements are non-zero. We term this as a *sparse system*. In terms of a difference equation, for example, a sparse system may be a high-dimensional system with only a few non-zero coefficients or it may be a system with an impulse response that is long but contains only a few non-zero terms. Multipath propagation [2], [3], sparse channel estimation [4], topology identification of interconnected systems [5], [6] and sparse initial state estimation [7] are examples involving systems that are high-order in terms of their ambient dimension but have a sparse (low-order) representation.

Inspired by the emerging field of Compressive Sensing (CS) [8], [9], in this paper, we aim at performing system identification of sparse systems using a number of observations that is smaller than the ambient dimension. We call this problem Compressive System Identification (CSI). CSI is beneficial in applications when only a limited data set is available. Moreover, CSI can help solve the issue of under and over parameterization, which is a common problem in parametric system identification. The chosen model structure, in terms of model order and number of delays and so on, on one hand should be rich enough to represent the behavior of the system and on the other hand should involve a minimal set of unknown parameters to minimize the variance of the parameter estimates. Under and over parameterization may have a considerable impact on the identification result, and choosing an optimal model structure is one of the primary challenges in system identification. Specifically, this can be a more problematic issue when: 1) the actual system to be identified is sparse and/or 2) it is a multivariable (Multi-Input Single-Output (MISO) or Multi-Input Multi-Output (MIMO)) system with I/O channels of different system orders and unknown (possibly large) input delays. Finding an optimal choice of the model structure for such systems is less likely to happen from cross-validation approaches.

Related works include regularization techniques such as the Least Absolute Shrinkage and Selection Operator

⋆B. M. Sanandaji, T. L. Vincent, and M. B. Wakin are with Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA. {bmolazem, tvincent, mwakin}@mines.edu.

†R. Tóth is with the Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands. r.toth@tudelft.nl.

⋄K. Poolla is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering, University of California, Berkeley, CA 94720, USA. poolla@berkeley.edu.

This work was partially supported by AFOSR Grant FA9550-09-1-0465, NSF Grant CCF-0830320, NSF Grant CNS-0931748 and NWO Grant 680-50-0927.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

(LASSO) algorithm [10] and the Non-Negative Garrote (NNG) method [11]. These methods were first introduced for linear regression models in statistics. There also exist some results on the application of these methods to Linear Time-Invariant (LTI) Auto Regressive with eXternal input (ARX) identification [12]. However, most of these results concern the stochastic properties of the parameter estimates in an asymptotic sense, with few results considering the limited data case. There is also some recent work on regularization of ARX parameters for LTV systems [13], [14].

In this paper, we consider CSI of ARX models for both LTI and LTV systems. We examine parameter estimation in the context of CS and formulate the identification problem as recovery of a *block-sparse* signal from an underdetermined set of linear equations. We discuss required measurements in terms of recovery conditions, derive bounds for such guarantees, and support our approach with simulations.

## II. Notation

In this section, we establish our notation. An LTI Single-Input Single-Output (SISO) ARX model [1] with parameters $\{n, m, d\}$ is given by the difference equation

$$\begin{aligned} y(t) + a_1 y(t-1) + \cdots + a_n y(t-n) = \\ b_1 u(t-d-1) + \cdots + b_m u(t-d-m) + e(t), \end{aligned} \quad (1)$$

where $y(t) \in \mathbb{R}$ is the output at time instant $t$, $u(t) \in \mathbb{R}$ is the input, $d$ is the input delay, and $e(t)$ is a zero mean stochastic noise process. Assuming $d+m \leq p$, where $p$ is the input maximum length (including delays), (1) can be written compactly as

$$y(t) = \boldsymbol{\phi}^T(t)\boldsymbol{\theta} + e(t) \quad (2)$$

where

$$\boldsymbol{\phi}(t) = \begin{bmatrix} -y(t-1) \\ \vdots \\ -y(t-n) \\ u(t-1) \\ \vdots \\ u(t-d-1) \\ \vdots \\ u(t-d-m) \\ \vdots \\ u(t-p) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ 0 \\ \vdots \\ b_1 \\ \vdots \\ b_m \\ \vdots \\ 0 \end{bmatrix},$$

$\boldsymbol{\phi}(t) \in \mathbb{R}^{n+p}$ is the data vector containing input-output measurements, and $\boldsymbol{\theta} \in \mathbb{R}^{n+p}$ is the parameter vector. The goal of the system identification problem is to estimate the parameter vector $\boldsymbol{\theta}$ from $M$ observations of the system. Taking $M$ consecutive measurements and putting them in a regression form, we have

$$\underbrace{\begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+M-1) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \boldsymbol{\phi}^T(t) \\ \boldsymbol{\phi}^T(t+1) \\ \vdots \\ \boldsymbol{\phi}^T(t+M-1) \end{bmatrix}}_{\Phi} \boldsymbol{\theta} + \underbrace{\begin{bmatrix} e(t) \\ e(t+1) \\ \vdots \\ e(t+M-1) \end{bmatrix}}_{\mathbf{e}}$$

or equivalently

$$\mathbf{y} = \Phi\boldsymbol{\theta} + \mathbf{e}. \quad (3)$$

In a noiseless scenario ($\mathbf{e} = \mathbf{0}$), from standard arguments in linear algebra, $\boldsymbol{\theta}$ can be exactly recovered from $M > n+p$ observations under the assumption of a persistently exciting input. Note that $\Phi$ in (3) is a concatenation of 2 blocks

$$\Phi = [\Phi_y | \Phi_u] \quad (4)$$

where $\Phi_y \in \mathbb{R}^{M \times n}$ and $\Phi_u \in \mathbb{R}^{M \times p}$ are Toeplitz matrices. Equation (4) can be extended for MISO systems with $l$ inputs as

$$\Phi = [\Phi_y | \Phi_{u_1} | \Phi_{u_2} | \cdots | \Phi_{u_l}] \quad (5)$$

where the $\Phi_{u_i}$'s are Toeplitz matrices, each containing regression over one of the inputs. The ARX model in (1) can be also represented as

$$\mathcal{A}(q^{-1})y(t) = q^{-d}\mathcal{B}(q^{-1})u(t) \quad (6)$$

where $q^{-1}$ is the backward time-shift operator, e.g., $q^{-1}y(t) = y(t-1)$, and $\mathcal{A}(q^{-1})$ and $\mathcal{B}(q^{-1})$ are vector polynomials defined as $\mathcal{A}(q^{-1}) = [1 \ a_1 q^{-1} \ \cdots \ a_n q^{-n}]$, and $\mathcal{B}(q^{-1}) = [b_1 q^{-1} \ \cdots \ b_m q^{-m}]$. For a MISO system with $l$ inputs, (6) extends to

$$\begin{aligned} \mathcal{A}(q^{-1})y(t) = \\ q^{-d_1}\mathcal{B}_1(q^{-1})u_1(t) + \cdots + q^{-d_l}\mathcal{B}_l(q^{-1})u_l(t) \end{aligned} \quad (7)$$

where $\mathcal{B}_i(q^{-1}), i = 1, 2, \cdots, l$, are low-order polynomials.

## III. CS Background and Recovery Algorithm

First introduced by Candès, Romberg and Tao [8], and Donoho [9], CS has emerged as a powerful paradigm in signal processing which enables the recovery of an unknown vector from an underdetermined set of measurements under the assumption of sparsity of the signal and certain conditions on the measurement matrix. The CS recovery problem can be viewed as recovery of a $K$-sparse signal $\mathbf{x} \in \mathbb{R}^N$ from its observations $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^M$ where $A \in \mathbb{R}^{M \times N}$ is the measurement matrix with $M < N$ (in many cases $M \ll N$). A $K$-sparse signal $\mathbf{x} \in \mathbb{R}^N$ is a signal of length $N$ with $K$ non-zero entries where $K < N$. The notation $K := \|\mathbf{x}\|_0$ denotes the sparsity level of $\mathbf{x}$. Since the null space of $A$ is non-trivial, there are infinitely many candidate solutions to the equation $\mathbf{b} = A\mathbf{x}$; however, it has been shown that under certain conditions on the measurement matrix $A$, CS recovery algorithms can recover that unique solution if it is suitably sparse.

Several recovery guarantees have been proposed in the CS literature. The Restricted Isometry Property (RIP) [15], the Exact Recovery Condition (ERC) [16], and mutual coherence [17], [18] are among the most important conditions. In this paper, our focus is on the mutual coherence due to its ease of calculation as compared to other conditions which are usually hard or even impossible to calculate. On the other hand, mutual coherence is a conservative measure as it only reflects the worst correlations in the matrix.

**Algorithm 1** The BOMP – block-sparse recovery

```
Require: matrix $A$, measurements $\mathbf{b}$, block size $n$, stopping criteria
Ensure: $\mathbf{r}^0 = \mathbf{b}$, $\mathbf{x}^0 = \mathbf{0}$, $\Lambda^0 = \emptyset$, $l = 0$
  repeat
      1. match: $\mathbf{e}_i = A_i^T \mathbf{r}^l, \quad i = 1, 2, \cdots, P$
      2. identify support: $\lambda = \arg\max_i \|\mathbf{e}_i\|_2$
      3. update the support: $\Lambda^{l+1} = \Lambda^l \cup \lambda$
      4. update signal estimate:
         $\mathbf{x}^{l+1} = \arg\min_{\mathbf{z}:\mathrm{supp}(\mathbf{z}) \subseteq \Lambda^{l+1}} \|\mathbf{b} - A\mathbf{z}\|_2$,
         where supp($\mathbf{z}$) indicates the blocks
         on which $\mathbf{z}$ is non-zero
      5. update residual estimate: $\mathbf{r}^{l+1} = \mathbf{b} - A\mathbf{x}^{l+1}$
      6. increase index $l$ by $1$
  until stopping criteria true
  output: $\widehat{\mathbf{x}} = \mathbf{x}^l$
```

*Definition 1 ( [17], [18]):* For a given matrix $A$, the *mutual coherence* equals the maximum normalized absolute inner product between two distinct columns of $A$, i.e,

$$\mu(A) := \frac{\max_{i,j \neq i} |\mathbf{a}_i^T \mathbf{a}_j|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}, \tag{8}$$

where $\{\mathbf{a}_i\}_{i=1}^N$ are the columns of $A$.

In general, CS recovery algorithms can be classified into two main types: 1) greedy algorithms such as Orthogonal Matching Pursuit (OMP) [18] and 2) convex optimization algorithms such as Basis Pursuit (BP) [19]. It has been shown [18] that the OMP and BP algorithms will recover any $K$-sparse signal $\mathbf{x} \in \mathbb{R}^N$ from $\mathbf{b} = A\mathbf{x}$ whenever

$$\mu(A) < \frac{1}{2K - 1}. \tag{9}$$

In other words, a smaller coherence indicates recovery of signals with more non-zero elements.

The sparse signals that are of our interest in this paper have a *block-sparse* structure, meaning that their non-zero entries appear in block locations.

*Definition 2:* Consider $\mathbf{x} \in \mathbb{R}^N$ as a concatenation of $P$ vector-blocks $\mathbf{x}_i \in \mathbb{R}^n$ where $N = Pn$ i.e.,

$$\mathbf{x} = [\mathbf{x}_1^T \cdots \mathbf{x}_i^T \cdots \mathbf{x}_P^T]^T. \tag{10}$$

A signal $\mathbf{x}$ is called block $K$-sparse if it has $K < P$ non-zero blocks.

There exist a few recovery algorithms that are adapted to recover such signals. In this paper, we focus on recovery of block-sparse signals via a greedy algorithm called Block Orthogonal Matching Pursuit (BOMP) [20]–[22]. We consider BOMP due to its ease of implementation and its flexibility in recovering block-sparse signals of different sparsity levels. The formal steps of the BOMP algorithm are listed in Algorithm 1 which finds a block-sparse solution to the equation $\mathbf{b} = A\mathbf{x}$.

The basic intuition behind BOMP is as follows. Due to the block sparsity of $\mathbf{x}$, the vector of observations $\mathbf{b}$ can be written as a succinct linear combination of the columns of $A$, with the selections of columns occurring in clusters due to the block structure of the sparsity pattern in $\mathbf{x}$. BOMP attempts to identify the participating indices by correlating the measurements $\mathbf{b}$ against the columns of $A$ and comparing the correlation statistics among different blocks. Once a significant block has been identified, its influence is removed from the measurements $\mathbf{b}$ via an orthogonal projection, and the correlation statistics are recomputed for the remaining blocks. This process repeats until convergence. Eldar et al. [22] proposed a sufficient condition for BOMP to recover any sufficiently concise block-sparse signal $\mathbf{x}$ from compressive measurements. This condition depends on two coherence metrics, the block and sub-block coherence of the matrix $A$. For a detailed description of these metrics see [22]. These metrics are basically related to the mutual coherence as defined in (8), although more adapted for block structure.

## IV. CSI OF LTI ARX MODELS

Identification of LTI ARX models in both SISO and MISO cases is considered in this section. As a first step towards CSI and for the sake of simplicity we consider the noiseless case. Inspired by CS, we show that in cases where the LTI system has a *sparse* impulse response, simultaneous ARX model identification and input delay estimation is possible from a small number of observations, even though this leaves the aforementioned linear equations highly underdetermined. We discuss the required number of measurements in terms of metrics that guarantee exact identification, derive bounds on such metrics, and suggest a pre-filtering scheme by which these metrics can be reduced. The companion paper [23] analyzes the consistency properties of CSI and explores the connection with LASSO and NNG sparse estimators.

### A. CSI of LTI Systems with Unknown Input Delays

Input delay estimation can be challenging, especially for large-scale multivariable (MISO or MIMO) systems when there exist several inputs with different unknown (possibly large) delays. Identification of such systems requires estimating (or guessing) the proper value of the $d_i$'s separately. Typically this is done via model complexity metrics such as the AIC or BIC, or via cross-validation by splitting the available data into an identification set and a validation set and estimating the parameters on the identification set for a fixed set of parameters $\{d_i\}$. This procedure continues by fixing another set of parameters, and finishes by selecting the parameters that give the best fit on the validation set. However, complete delay estimation would require estimation and cross validation with all possible delay combinations, which can grow quickly with the number of inputs. For instance, with 5 inputs, checking for delays in each channel between 1 and 10 samples requires solving $10^5$ least-squares problems. A review of other time-delay estimation techniques is given in [24]. For a sufficiently large number of inputs with possibly large delays, we will show that by using the tools in CS, it is possible to implicitly estimate the delays by favoring block-sparse solutions for $\boldsymbol{\theta}$.

Letting $m_i$ be the length of $\mathcal{B}_i$ and bounding the maximum length (including delays) for all inputs by $p$ ($\max_i(d_i + m_i) \leq p$), we build the regression matrix with each $\Phi_{u_i} \in \mathbb{R}^{M \times p}$ to be a Toeplitz matrix associated with one input. This results in an $M \times (n + lp)$ matrix $\Phi$. However, considering a low-order polynomial for each input ($\max_i m_i \leq m$) for some $m$, the corresponding parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{n+lp}$ has at most $n + lm$ non-zero entries. Assuming $m < p$, this formulation suggests *sparsity* of the parameter vector $\boldsymbol{\theta}$ and encourages us to use the tools in CS for recovery. Moreover, this allows us to do the identification from an underdetermined set of equations $\Phi$ where $M < n + lp$.

### *B. Simulation Results*

Fig. 1(a) illustrates the recovery of a $\{2, 2, 40\}$ SISO LTI ARX model where $m$ and $d$ are unknown. The only knowledge is of $p = 62$. For each system realization, the input is generated as an independent and identically distributed (i.i.d.) Gaussian random sequence. Assuming at least $d$ iterations of the simulation have passed, $M$ consecutive samples of the output are taken. As $n$ is known, we modify the BOMP algorithm to include the first $n$ locations as part of the support of $\boldsymbol{\theta}$. The plot shows the recovery success rate over 1000 realizations of the system. As shown in Fig. 1(a), with 25 measurements, the system is perfectly identified in 100% of the trials. The average coherence value as defined in (8) is also depicted in Fig. 1(b) (solid curve). After taking a certain number of measurements, the average coherence converges to a constant value (dashed line). We will address this in detail in the next section.

Identification of a MISO system is shown in Fig. 2 where the actual system has parameters $n = 2$, $m = 2$ for all inputs, and $d_1 = 60, d_2 = 21, d_3 = 10, d_4 = 41$. Assuming $p = 64$, the parameter vector $\boldsymbol{\theta}$ has 258 entries, only 10 of which are non-zero. Applying the BOMP algorithm with $n$ given and $m$ and $\{d_i\}$ unknown, implicit input delay estimation and parameter identification is possible in 100% of the trials by taking $M = 150$ measurements.

### *C. Bound on Coherence*

As depicted in Fig. 1(b), the typical coherence $\mu(\Phi)$ has an asymptotic behavior. In this section, we derive a lower bound on the typical value of $\mu(\Phi)$ for SISO LTI ARX models. Specifically, for a given system excited by a random i.i.d. Gaussian input, we are interested in finding $\mathbf{E}\left[\mu(\Phi)\right]$ where $\mu$ is defined as in (8) and $\Phi$ is as in (3).

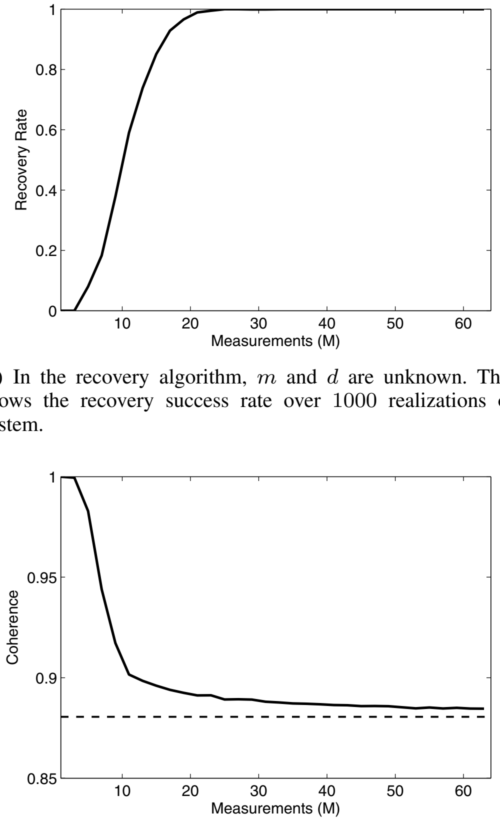
*Theorem 1:* Consider the system described by difference equation in (1) (ARX model $\{n, m, d\}$) is characterized by its impulse response $h(k)$ in a convolution form as

$$y(t) = \sum_{k=-\infty}^{\infty} h(k) u(t-k). \tag{11}$$

Then, for a zero mean, unit variance i.i.d. Gaussian input,

$$\lim_{M \to \infty} \mathbf{E}\left[\mu(\Phi)\right] \geq \max_{s \neq 0} \left\{ \frac{|\mathcal{H}(s)|}{\|h\|_2^2}, \frac{|h(s)|}{\|h\|_2} \right\} \tag{12}$$

where $\mathcal{H}(s) = \sum_{k=-\infty}^{\infty} h(k) h(k+s)$.



(a) In the recovery algorithm, $m$ and $d$ are unknown. The plot shows the recovery success rate over 1000 realizations of the system.

(b) Averaged mutual coherence of $\Phi$ over 1000 realizations of the system (solid curve). Lower bound of Theorem 1 (dashed line).

Fig. 1. CSI results on a $\{2, 2, 40\}$ SISO LTI ARX system.

*Proof:* See Appendix $A$. ■

*Discussion*: As Theorem 1 suggests, the typical coherence of $\Phi$ is bounded below by a non-zero value that depends on the impulse response of the system and it has an asymptotic behavior. For example, for the system given in Fig. 1, the typical coherence does not get lower than 0.88 even for large $M$. With this value of coherence, the analytical recovery guarantees for the BOMP algorithm [22], which can be reasonably represented by mutual coherence defined in (8), do not guarantee recovery of any one-block sparse signals. However, as can be seen in Fig. 1(a), perfect recovery is possible. This indicates a gap between available analytical guarantee and the true recovery performance for ARX systems. This suggests that coherence-based performance guarantees for matrices that appear in ARX identification are not sharp tools as they only reflect the worst correlations in the matrix. As a first step towards investigating this gap, we suggest a pre-filtering scheme by which the coherence of such matrices can be reduced.

### *D. Reducing Coherence by Pre-Filtering*

In this section, we show that we can reduce the coherence by designing a pre-filter $g$ applied on $u$ and $y$.

*Theorem 2:* Assume the system described as in Theorem 1. Given a filter $g$, define $u_g = u * g$ and $y_g = y * g$.

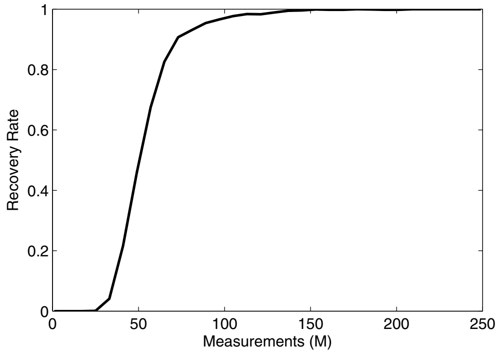
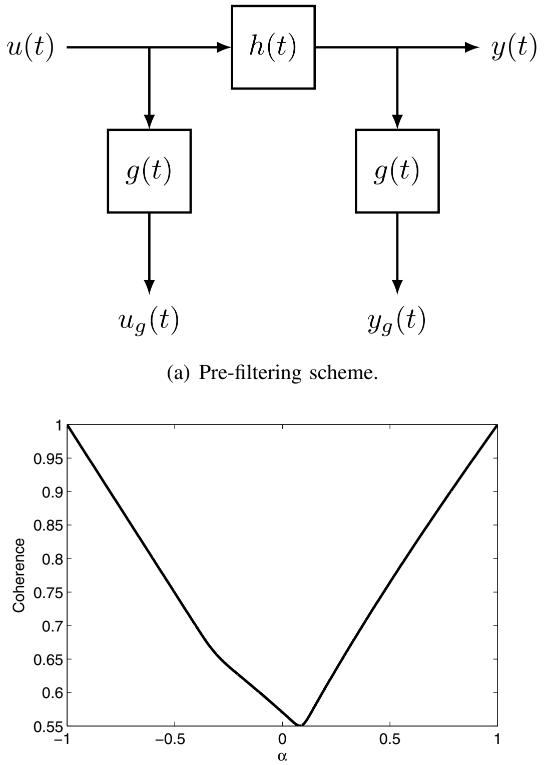


Fig. 2. CSI results on a $\{2, 2, \{60, 21, 10, 41\}\}$ MISO LTI ARX system. In the recovery algorithm, $m$ and $\{d_i\}_{i=1}^{4}$ are unknown. The plot shows the recovery success rate over 1000 realizations of the system.



(a) Pre-filtering scheme.

(b) For each $\alpha$, the filter $G(z)$ is applied on the input/output signals and the limit of the expected value of coherence is calculated over 1000 realizations of system.

Fig. 3. Reducing coherence by pre-filtering.

Build the regression matrix $\Phi_g$ from $u_g$ and $y_g$ as in (3). The pre-filtering scheme is shown in Fig. 3(a). Then we have

$$\lim_{M\to\infty} \mathbf{E}\left[\mu(\Phi_g)\right] \geq \max_{s\neq 0}\left\{\frac{|\mathcal{G}(s)|}{\|g\|_2^2}, \frac{|\mathcal{F}(s)|}{\|f\|_2^2}, \frac{|\mathcal{GF}(s)|}{\|g\|_2\|f\|_2}\right\}$$

where $f = g * h$, $\mathcal{G}(s) = \sum_{k=-\infty}^{\infty} g(k)g(k+s)$, $\mathcal{F}(s) = \sum_{k=-\infty}^{\infty} f(k)f(k+s)$, and $\mathcal{GF}(s) = \sum_{k=-\infty}^{\infty} g(k)f(k+s)$.

*Proof:* See Appendix $B$. ■

Theorem 2 suggests that by choosing an appropriate filter $g(t)$, the typical coherence can possibly be reduced, although it is bounded below by a non-zero value. We follow the discussion by showing how the coherence of $\Phi$ can be reduced by pre-filtering within an illustrative example. Consider a SISO system characterized by the transfer function

$$H(z) = \frac{z - 0.4}{(z+0.9)(z+0.2)}. \tag{13}$$

Using the bound given in Theorem 1, for large $M$, $\mathbf{E}\left[\mu\Phi\right] \geq 0.95$ which indicates a highly correlated matrix $\Phi$. However, using the analysis given in Theorem 2, we can design a filter $G(z)$ such that the coherence of the resulting matrix $\Phi_g$ is reduced almost by half. For example, consider a notch filter $G(z)$ given by

$$G(z) = \frac{z + 0.9}{(z + \alpha)} \tag{14}$$

where $\alpha$ is a parameter to be chosen. For a given $\alpha$, the filter $G(z)$ is applied on the input/output data as illustrated in Fig. 3(a) and the average coherence of $\Phi_g$ is calculated. The result of this pre-filtering and its effect on the coherence is shown in Fig. 3(b). The results indicate that actual performance of $\Phi$ may actually be better than what $\mu(\Phi)$ suggests. As it can be seen, for $\alpha$ around $0.1$, the coherence is reduced to $0.55$ which is almost half of the primary coherence.

## V. CSI of LTV ARX Models

In (2) the parameters are assumed to be fixed over time. In this section, we study ARX models where the parameter vector $\boldsymbol{\theta}(t)$ is varying over time. As an extension of (2), for time-varying systems, we have

$$y(t) = \boldsymbol{\phi}^T(t)\boldsymbol{\theta}(t) + e(t).$$

Collecting $M$ consecutive measurements of such a system and following similar steps, for a SISO LTV ARX model we can formulate the parameter estimation problem as

$$\underbrace{\begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+M-1) \end{bmatrix}}_{\mathbf{y}} =$$

$$\underbrace{\begin{bmatrix} \boldsymbol{\phi}^T(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\phi}^T(t+1) & \mathbf{0} & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\phi}^T(t+M-1) \end{bmatrix}}_{\Omega} \underbrace{\begin{bmatrix} \boldsymbol{\theta}(t) \\ \boldsymbol{\theta}(t+1) \\ \vdots \\ \boldsymbol{\theta}(t+M-1) \end{bmatrix}}_{\boldsymbol{\vartheta}} + \mathbf{e}$$

or equivalently

$$\mathbf{y} = \Omega\boldsymbol{\vartheta} + \mathbf{e} \tag{15}$$

where for simplicity $d = 0$, $p = m$, $\mathbf{y} \in \mathbb{R}^M$, $\Omega \in \mathbb{R}^{M\times M(n+m)}$ and $\boldsymbol{\vartheta} \in \mathbb{R}^{M(n+m)}$. The goal is to solve (15) for $\boldsymbol{\vartheta}$ from $\mathbf{y}$ and $\Omega$. Typical estimation is via

$$\min_{\boldsymbol{\vartheta}} \|\mathbf{y} - \Omega\boldsymbol{\vartheta}\|_2^2. \tag{16}$$

However, the minimization problem in (16) contains an underdetermined set of equations ($M < M(n+m)$) and therefore has many solutions.

### A. Piecewise-Constant $\boldsymbol{\theta}(t)$ and Block-Sparse Recovery

Assuming $\boldsymbol{\theta}(t)$ is piecewise-constant, we show how the LTV ARX identification can be formulated as recovery of a block-sparse signal. Using the developed tools in CS we show the identification of such systems can be done from relatively few measurements. Assume that $\mathbf{e} = \mathbf{0}$ and that $\boldsymbol{\theta}(t)$ changes only at a few time instants $t_i \in \mathcal{C}$ where $\mathcal{C} \triangleq \{t_1, t_2, \dots\}$ with $|\mathcal{C}| \ll M$, i.e.,

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t_i), \qquad t_i \le t < t_{i+1}. \tag{17}$$

Note that neither the change moments $t_i$ nor the number of changes is known a priori to the identification algorithm. An example of $\boldsymbol{\vartheta}$ would be

$$\boldsymbol{\vartheta} = \begin{bmatrix} \boldsymbol{\theta}^T(t_1) & \cdots & \boldsymbol{\theta}^T(t_1) & \boldsymbol{\theta}^T(t_2) & \cdots & \boldsymbol{\theta}^T(t_2) \end{bmatrix}^T \tag{18}$$

which has 2 different constant pieces, i.e., $\mathcal{C} = \{t_1, t_2\}$. In order to exploit the existing sparsity pattern in $\boldsymbol{\vartheta}$, define the differencing operator

$$\Delta = \begin{bmatrix} -I_{n+m} & 0_{n+m} & \cdots & \cdots & 0_{n+m} \\ I_{n+m} & -I_{n+m} & \ddots & \ddots & \vdots \\ 0_{n+m} & I_{n+m} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0_{n+m} \\ 0_{n+m} & \cdots & 0_{n+m} & I_{n+m} & -I_{n+m} \end{bmatrix}.$$

Applying $\Delta$ to $\boldsymbol{\vartheta}$, we define $\boldsymbol{\vartheta}_\delta$ as

$$\boldsymbol{\vartheta}_\delta = \Delta \boldsymbol{\vartheta}, \tag{19}$$

which has a block-sparse structure. For the given example in (18), we have

$$\boldsymbol{\vartheta}_\delta = \begin{bmatrix} -\boldsymbol{\theta}^T(t_1)\, \mathbf{0} \cdots \mathbf{0}\, \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2)\, \mathbf{0} \cdots \mathbf{0} \end{bmatrix}^T. \tag{20}$$

The vector $\boldsymbol{\vartheta}_\delta \in \mathbb{R}^{M(n+m)}$ in (20) now has a block-sparse structure: out of its $M(n+m)$ entries, grouped in $M$ blocks of length $n+m$, only a few of them are non-zero and they appear in block locations. The number of non-zero blocks corresponds to the number of different levels of $\boldsymbol{\theta}(t)$. In the example given in (18), $\boldsymbol{\theta}(t)$ takes 2 different levels over time and thus, $\boldsymbol{\vartheta}_\delta$ has a block-sparsity level of 2 with each block size of $n+m$. By this formulation, the parameter estimation of LTV ARX models with piecewise-constant parameter changes can be cast as recovering a block-sparse signal $\boldsymbol{\vartheta}_\delta$ from measurements

$$\mathbf{y} = \Omega_\delta \boldsymbol{\vartheta}_\delta \tag{21}$$

where $\Omega_\delta = \Omega \Delta^{-1}$.

### B. Identifiability Issue

Before presenting the simulation results, we address the identifiability issue faced in the LTV case. The matrix $\Omega_\delta$ has the following structure.

$$\Omega_\delta = \begin{bmatrix} -\boldsymbol{\phi}^T(t) & \mathbf{0} & \mathbf{0} & \cdots \\ -\boldsymbol{\phi}^T(t+1) & -\boldsymbol{\phi}^T(t+1) & \mathbf{0} & \cdots \\ -\boldsymbol{\phi}^T(t+2) & -\boldsymbol{\phi}^T(t+2) & -\boldsymbol{\phi}^T(t+2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

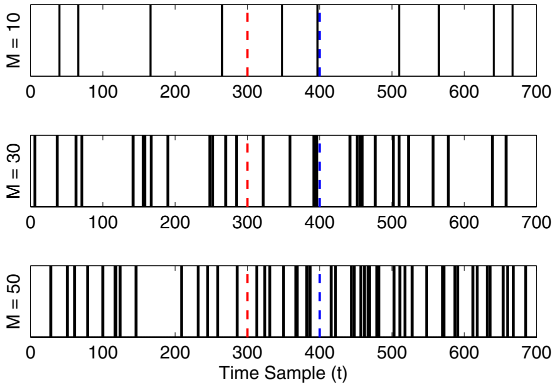


Fig. 4. Random sampling scheme for $M = \{10, 30, 50\}$ measurements. Samples are chosen randomly according to a uniform distribution. System parameters are assumed to change at $t = 300$ and $t = 400$.

If the change in the system actually happens at time instant $t+2$, the corresponding solution to (21) has the form

$$\boldsymbol{\vartheta}_\delta = \begin{bmatrix} -\boldsymbol{\theta}^T(t_1) & \mathbf{0} & \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2) & \mathbf{0} & \cdots \end{bmatrix}^T.$$

However, due to the special structure of the matrix $\Omega_\delta$, there exist other solutions to this problem. For example

$$\widehat{\boldsymbol{\vartheta}}_\delta = \begin{bmatrix} -\boldsymbol{\theta}^T(t_1) & \mathbf{0} & \boldsymbol{\theta}^T(t_1) - \boldsymbol{\theta}^T(t_2) + \boldsymbol{\gamma}^T & -\boldsymbol{\gamma}^T & \cdots \end{bmatrix}^T$$

is another solution where $\boldsymbol{\gamma}$ is a vector in the null space of $\boldsymbol{\phi}^T(t)$, i.e., $\boldsymbol{\phi}^T(t)\boldsymbol{\gamma} = 0$. However, this only results in a small ambiguity in the solution around the transition point. Therefore, $\widehat{\boldsymbol{\vartheta}}_\delta$ can be considered as an acceptable solution as $\widehat{\boldsymbol{\vartheta}} = \Delta^{-1}\widehat{\boldsymbol{\vartheta}}_\delta$ is exactly equal to the true parameter vector $\boldsymbol{\vartheta}$ except at very few time instants around the transition point. In the next section, we consider $\widehat{\boldsymbol{\vartheta}}_\delta$ as a valid solution.

### C. Sampling Approach for LTV System Identification

In this section, we suggest a sampling scheme for identifying LTV systems. Note that in a noiseless scenario, the LTV identification can be performed by taking consecutive observations in a frame, identifying the system on that frame, and then moving the frame forward until we identify a change in the system. Of course, this can be very inefficient when the time instants at which the changes happen are unknown to us beforehand as we end up taking many unnecessary measurements. As an alternative, we suggest a *random* sampling scheme (as compared to consecutive sampling) for identifying such LTV systems. Fig. 4 shows examples of this sampling approach for $M = 10$, $M = 30$ and $M = 50$ measurements. As can be seen, the samples are chosen randomly according to a uniform distribution. Note that these samples are not necessarily consecutive. By this approach, we can dramatically reduce the required number of measurements for LTV system identification.

### D. Simulation Results

Consider a system described by its $\{2, 2, 0\}$ ARX model

$$y(t) + a_1 y(t-1) + a_2 y(t-2) = b_1 u(t-1) + b_2 u(t-2) \tag{22}$$

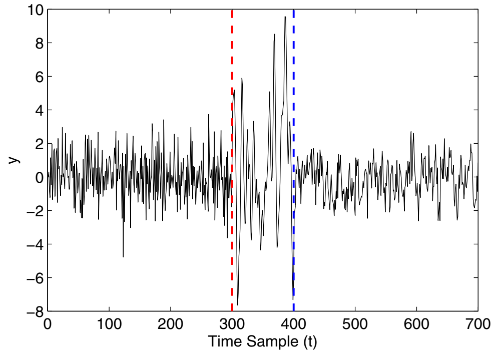


Fig. 5. Output of a 3-model system. System parameters change at $t = 300$ and $t = 400$. At the time of change, all the system parameters change.

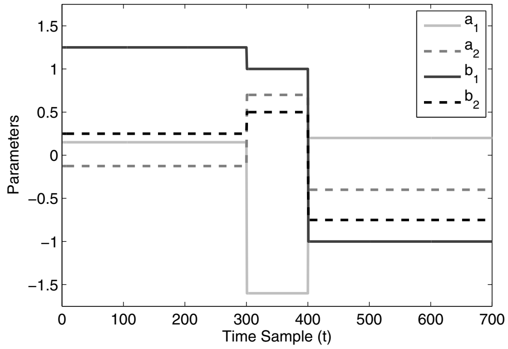


Fig. 6. Time-varying parameters of a 3-model system.

with i.i.d. Gaussian input $u(t) \sim \mathcal{N}(0, 1)$. Fig. 5 shows one realization of the output of this system whose parameters are changing over time as shown in Fig. 6. As can be seen, the parameters change in a piecewise-constant manner over 700 time instants at $t = 300$ and $t = 400$. The goal of the identification is to identify the parameters of this time-variant system along with the location of the changes.

Fig. 7 illustrates the recovery performance of 4 LTV systems, each with a different number of changes over time. For each measurement sequence (randomly selected), 1000 realizations of the system are carried out. We highlight two points about this plot. First, we are able to identify a system (up to the ambiguity around the time of change as discussed in Section V-B) which changes 3 times over 700 time instants by taking only 50 measurements without knowing the location of the changes. Second, the required number of measurements for perfect recovery scales with number of changes that a system undergoes over the course of identification. Systems with more changes require more measurements to be identified.

## VI. Conclusion

We considered CSI of LTI and LTV ARX models for systems with limited data sets. We showed in cases where the LTI system has possibly many inputs with different unknown delays, simultaneous ARX model identification and input delay estimation is possible from a small number of observations. We also considered identifying LTV ARX models. In particular, we considered systems with parameters changing only at a few time instants where neither the change moments nor the number of changes is known a priori. The main technical novelty of our approach is in casting the identification problem in the context of CS as recovery of a *block-sparse* signal from an underdetermined set of linear equations. We discussed the required number of measurements in terms of recovery conditions and derived bounds for such guarantees and supported our approach by illustrative simulations.

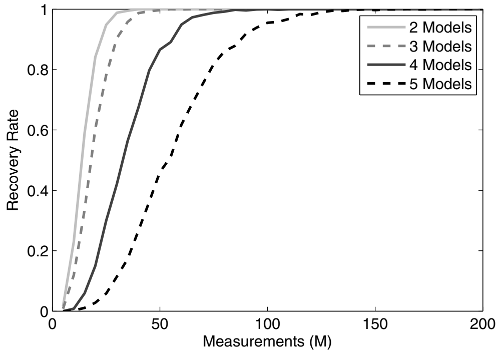


Fig. 7. Recovery performance of 4 different systems. The plots show the recovery success rate over 1000 realizations of the system.

## Appendix

### A. Proof of Theorem 1

*Proof:* Without loss of generality, assume $d = 0$ as the input delays do not affect the coherence of $\Phi$. Using the definition of $\mu(\Phi)$, we can write $\mu(\Phi) = \|\boldsymbol{\mu}_\Phi\|_\infty$ where $\boldsymbol{\mu}_\Phi$ is a vector whose entries are all the normalized distinct inner products of the columns of $\Phi$ and $\|\cdot\|_\infty$ is the maximum absolute entry of a vector. From Jensen's inequality for convex functions ($\|\cdot\|_\infty$), we have

$$\mathbf{E}\left[\mu(\Phi)\right] = \mathbf{E}\left[\|\boldsymbol{\mu}_\Phi\|_\infty\right] \geq \|\mathbf{E}\left[\boldsymbol{\mu}_\Phi\right]\|_\infty.$$

First we look at the numerator of the entries of $\boldsymbol{\mu}_\Phi$. From the definition of $\Phi$, $\forall \boldsymbol{\phi}_i, \boldsymbol{\phi}_{i+s} \in \Phi_y$, $s \neq 0$,

$$\boldsymbol{\phi}_i^T \boldsymbol{\phi}_{i+s} = \sum_{t=t_0}^{t_0+M} y(t)y(t-s). \tag{23}$$

Combining (23) with (11) and reordering the sums we have

$$\begin{aligned}\boldsymbol{\phi}_i^T \boldsymbol{\phi}_{i+s} &= \sum_{t=t_0}^{t_0+M} y(t)y(t-s) = \\ \sum_{t=t_0}^{t_0+M} &\left(\sum_{k=-\infty}^{\infty} h(k)u(t-k)\right)\left(\sum_{l=-\infty}^{\infty} h(l)u(t-l-s)\right) = \\ &\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h(k)h(l)\sum_{t=t_0}^{t_0+M} u(t-k)u(t-l-s).\end{aligned} \tag{24}$$

Taking the expected value of both sides of (24), we have

$$\mathbf{E}\left[\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i+s}\right] = M\sum_{l=-\infty}^{\infty} h(l)h(l+s) \tag{25}$$

where we used the fact that $\mathbf{E}\left[u(t-k)u(t-l-s)\right]=1$ for $k=l+s$ and 0 otherwise. Similarly, $\forall\boldsymbol{\phi}_i\in\Phi_y, \forall\boldsymbol{\phi}_{i+s}\in\Phi_u$,

$$\begin{aligned}\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i+s} = \sum_{t=t_0}^{t_0+M} y(t)u(t-s) = \\ \sum_{t=t_0}^{t_0+M}\left(\sum_{l=-\infty}^{\infty} h(l)u(t-l)\right)u(t-s) = \\ \sum_{l=-\infty}^{\infty} h(l)\sum_{t=t_0}^{t_0+M} u(t-l)u(t-s).\end{aligned} \tag{26}$$

Taking the expected value of both sides of (26), we have

$$\mathbf{E}\left[\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i+s}\right] = Mh(s). \tag{27}$$

It is trivial to see that $\forall\boldsymbol{\phi}_i,\boldsymbol{\phi}_{i+s}\in\Phi_u$ with $s\neq 0$, $\mathbf{E}\left[\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i+s}\right]=0$. Using concentration of measure inequalities, it can be shown that as $M\to\infty$, the entries of the denominator of $\boldsymbol{\mu}_\Phi$ are highly concentrated around their expected value [3]. We have $\forall\boldsymbol{\phi}_i\in\Phi_u, \mathbf{E}\left[\|\boldsymbol{\phi}_i\|_2^2\right]=M$ and $\forall\boldsymbol{\phi}_i\in\Phi_y$, $\mathbf{E}\left[\|\boldsymbol{\phi}_i\|_2^2\right]=M\|h\|_2^2$. By putting together (25) and (27) and applying the required column normalizations the proof is complete. ■

### *B. Proof of Theorem 2*

*Proof:* We follow a similar argument to the proof of Theorem 1. Define $u_g(t)=\sum_{k=-\infty}^{\infty} g(k)u(t-k)$ and $y_g(t)=\sum_{k=-\infty}^{\infty} g(k)y(t-k)$. Then we have

$$\begin{aligned}\sum_{t=t_0}^{t_0+M} y_g(t)u_g(t-s) = \\ \sum_{t=t_0}^{t_0+M}\left(\sum_{k=-\infty}^{\infty} g(k)y(t-k)\right)\left(\sum_{l=-\infty}^{\infty} g(l)u(t-l-s)\right)\end{aligned}$$

and by taking the expected value of both sides, we get

$$\mathbf{E}\left[\sum_{t=t_0}^{t_0+M} y_g(t)u_g(t-s)\right] = M\sum_{l=-\infty}^{\infty} g(l)f(l+s)$$

where $f=g*h$. In a similar way,

$$\mathbf{E}\left[\sum_{t=t_0}^{t_0+M} y_g(t)y_g(t-s)\right] = M\sum_{k=-\infty}^{\infty} f(k)f(k+s),$$

$$\mathbf{E}\left[\sum_{t=t_0}^{t_0+M} u_g(t)y_g(t-s)\right] = M\sum_{k=-\infty}^{\infty} f(k)g(k+s),$$

$$\mathbf{E}\left[\sum_{t=t_0}^{t_0+M} u_g(t)u_g(t-s)\right] = M\sum_{k=-\infty}^{\infty} g(k)g(k+s).$$

■

## References

[1] L. Ljung, *System Identification - Theory for the User*. Prentice-Hall, 2nd edition, 1999.

[2] J. Romberg, "Compressive sensing by random convolution," *SIAM Journal on Imaging Sciences*, vol. 2, no. 4, pp. 1098–1128, 2009.

[3] 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," *Proceedings of the* 49*th IEEE Conference on Decision and Control*, pp. 2922–2929, 2010.

[4] J. Haupt, W. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5862–5875, 2010.

[5] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Exact topology identification of large-scale interconnected dynamical systems from compressive observations," *Proceedings of the* 2011 *American Control Conference*, pp. 649–656, 2011.

[6] ——, "Compressive topology identification of interconnected dynamic systems via clustered orthogonal matching pursuit," *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference*, 2011.

[7] M. B. Wakin, B. M. Sanandaji, and T. L. Vincent, "On the observability of linear systems from random, compressive measurements," *Proceedings of the* 49*th IEEE Conference on Decision and Control*, pp. 4447–4454, 2010.

[8] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[9] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[10] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[11] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[12] C. Lyzell, J. Roll, and L. Ljung, "The use of nonnegative garrote for order selection of ARX models," *Proceedings of 47th IEEE Conference on Decision and Control*, pp. 1974–1979, 2008.

[13] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automatica*, vol. 46, no. 6, pp. 1107–1111, 2010.

[14] I. Maruta and T. Sugie, "A new approach for modeling hybrid systems based on the minimization of parameters transition in linear time-varying models," *Proceedings of the* 49*th IEEE Conference on Decision and Control*, pp. 117–1182, 2010.

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

[16] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1030–1051, 2006.

[17] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[18] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[19] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[20] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[21] ——, "Block-sparsity and sampling over a union of subspaces," *Proceedings of the 16th international conference on Digital Signal Processing*, pp. 1–8, 2009.

[22] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[23] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Compressive system identification in the linear time-invariant framework," *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference*, 2011.

[24] S. Bjorklund and L. Ljung, "A review of time-delay estimation techniques," *Proceedings of the* 42*th IEEE Conference on Decision and Control*, pp. 2502–2507, 2003.