

RLS SPARSE SYSTEM IDENTIFICATION USING LAR-BASED SITUATIONAL AWARENESS

Catia Valdman, Marcello L. R. de Campos

Federal University of Rio de Janeiro (UFRJ)
Program of Electrical Engineering
Rio de Janeiro, Brazil

cvaldman@inpi.gov.br, mcampos@ieee.org

José A. Apolinário Jr.

Military Institute of Engineering (IME)
Department of Electrical Engineering
Rio de Janeiro, Brazil

apolin@ime.eb.br

ABSTRACT

In this paper we propose the combination of the recursive least squares (RLS) and the least angle regression (LAR) algorithms for nonlinear system identification. In the application of interest, the model possesses a large number of coefficients, of which only few are different from zero. We use the LAR algorithm together with a geometrical stopping criterion to establish the number and position of the coefficients to be estimated by the RLS algorithm. The output error is used for indicating model inadequacy and therefore triggering the LAR algorithm. The proposed scheme is capable of modeling intrinsically sparse systems with better accuracy than the RLS algorithm alone, and lower energy consumption.

Index Terms— Nonlinear systems, Recursive Least Squares, Least Angle Regression, Volterra

1. INTRODUCTION

Increasing demand for systems capable of performing well with minimum power consumption has pushed the development of signal processing algorithms that exploit system sparsity. The Least Angle Regression (LAR) algorithm is a greedy stepwise algorithm that models heuristically sparse systems [1]. The algorithm has been applied successfully in several applications, as in [2–7]. Given a set of data, at each iteration the LAR algorithm adds a new non-zero coefficient to its active set, thus offering solutions ranging iteratively from highly sparse to completely dense [8]. Among other characteristics, the LAR algorithm provides a complete regularization path [1], which, if not stopped, leads to the LS solution; stopping this path at a certain point is the focus of this work.

In [6, 7], the LAR algorithm was used in tandem with a Volterra filter for the identification of nonlinear systems. Based on an ingenious technique for determining the number of nonzero coefficients [9], the algorithm was able to provide an estimator using only a subset of coefficients, which

were supposedly the most important ones. However, the LAR algorithm works with a batch of data and cannot update recursively its coefficients as more data is gathered. This limitation hinders its utilization in real-time applications, for it makes the algorithm prohibitively complex for large amounts of data. Overcoming the limitations imposed by the offline characteristics of the LAR algorithm, yet retaining its ability to indicate which coefficients matter most, would be most valuable for the case of sparse systems with a large number of coefficients.

We propose herein the use of the recursive least squares (RLS) algorithm for estimating the coefficients indicated by the LAR algorithm with a geometrical stopping criterion (which is referred to as GLAR). Output-error monitoring indicates model inadequacy, in which case the LAR algorithm together with the geometrical stopping criterion provides the information of the coefficients that need to be estimated. Although not changing peak complexity, we show that the proposed scheme offers a reduced time-averaged complexity, which could lead, in practical implementations, to an overall energy reduction.

Previous works related to sparse system identification (though not using the LAR algorithm) include [10–12]. In both articles, the RLS algorithm is applied with a model selection criterion, such as AIC [13] and BIC [14], to identify sparse systems. These selection criteria also offer the number of nonzero coefficients, but requires testing all orders prior to determining the desired value.

In [10] a convenient maximum number of nonzero coefficients is also needed for an improved performance. The LAR algorithm with a geometrical stopping criterion as seen herein is able to provide a number of nonzero coefficients with neither the need of estimating all coefficients nor a pre-established threshold.

The rest of this paper is organized as follows: Section 2 quickly reviews the LAR algorithm. The combination of LAR and RLS algorithms for situational awareness is presented in Section 3 and its application to nonlinear system identification is shown in Section 4. Conclusions are drawn in Section 5.

The authors want to express their gratitude to CAPES (Project Prodefesa, Process no. 23038.009094/2013-83), CNPq, FINEP, and INPI-BR for partial financial support of this work.

2. THE LAR ALGORITHM WITH THE GEOMETRICAL STOPPING CRITERION

The LAR algorithm was introduced in 2004 by Efron et al. [1] and has been extensively studied in the past ten years (e.g., [8, 15–20]). We will focus our discussion of the algorithm on the geometrical stopping criterion [9] for determining how many and which coefficients need to be estimated by the LAR algorithm.

The input data vector ($J \times 1$) is defined as (see Fig. 1.a)

$$\mathbf{x}(k) = [x_1(k) \ \cdots \ x_j(k) \ \cdots \ x_J(k)]^T, \quad (1)$$

where k is the time index, $k = 1, \dots, K$, and j is the channel index, $j = 1, \dots, J$. Gathering all input samples from $k = 1$ to $k = K$, the $(K \times J)$ input matrix is defined as $\mathbf{X} = [\mathbf{x}(1) \ \mathbf{x}(2) \ \cdots \ \mathbf{x}(K)]^T$. However, for the purposes of describing the LAR algorithm, an alternative representation may be preferable (see Fig. 1.b):

$$\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_j \ \cdots \ \mathbf{x}_J], \quad (2)$$

where $\mathbf{x}_j = [x_j(1) \ \cdots \ x_j(k) \ \cdots \ x_j(K)]^T$ carries the data of the j th channel, therefore presented to the j th coefficient, for all time instants, $k = 1, \dots, K$.

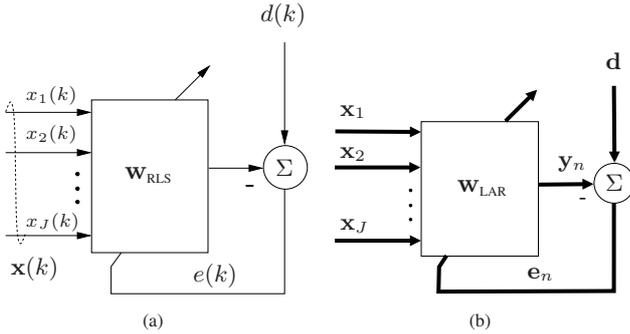


Fig. 1. (a) Block diagram for an adaptive filter operating recursively in time, k , and (b) block diagram for an algorithm operating with a batch of data.

Given a batch of data available in matrix \mathbf{X} , each iteration of the LAR algorithm seeks a new channel to be activated. Therefore, we might say that the n th iteration (in the LAR algorithm) is not related to time index k , but to the number of nonzero coefficients selected so far.

Let \mathbf{e}_n be the $(K \times 1)$ prediction error vector after n channels have been activated, i.e., at the n th iteration. We define the $(J \times 1)$ correlation vector \mathbf{c}_n between the input matrix and the prediction error vector as

$$\mathbf{c}_n = \mathbf{X}^T \mathbf{e}_n = [\mathbf{x}(1) \ \cdots \ \mathbf{x}(K)] \begin{bmatrix} e_n(1) \\ \vdots \\ e_n(K) \end{bmatrix}. \quad (3)$$

In the equation above, the j th element of vector \mathbf{c}_n , $c_{j,n}$, corresponds to an estimate of the correlation between the error (at iteration n) and the input signal in the j th channel.

We assume that the data for each channel has been normalized for zero mean and unitary length [1]. The j th correlation value in (3) can be rewritten as

$$c_{j,n} = \mathbf{x}_j^T \mathbf{e}_n = \underbrace{\|\mathbf{x}_j\|}_{=1} \|\mathbf{e}_n\| \cos \theta_{j,n} \quad (4)$$

such that $\theta_{j,n} = \arccos \frac{\mathbf{x}_j^T \mathbf{e}_n}{\|\mathbf{e}_n\|}$, where $\theta_{j,n}$ is the angle between the j th coefficient data vector and the error vector at iteration n .

In the last iteration ($n = J$), the LAR algorithm yields the LS solution [1]. Therefore, as stated by the orthogonality principle, $\theta_{j,J} = 90^\circ$, $j = 1, \dots, J$.

Let \mathcal{A}_n be the active set, i.e., the set of n indexes corresponding to channels which have been activated. The inactive set, $\bar{\mathcal{A}}_n$, is its complementary set and contains the indexes of coefficients which have been kept equal to zero. According to the LAR algorithm, the absolute values of the correlation coefficients for all channels in the active set are equal. Let this value be $C_{max,n}$, i.e.

$$|c_{j,n}| \begin{cases} = C_{max,n}, & j \in \mathcal{A}_n \\ < C_{max,n}, & j \in \bar{\mathcal{A}}_n \end{cases} \quad (5)$$

which can be simplified to $c_{j,n} < C_{max,n}$ when $j \in \bar{\mathcal{A}}_n$.

For the coefficients in the active set, $j \in \mathcal{A}_n$, the angle $\theta_{j,n}$ is given by

$$\theta_{j,n} = \begin{cases} \arccos \frac{-C_{max,n}}{\|\mathbf{e}_n\|}, & c_{j,n} < 0 \\ \arccos \frac{C_{max,n}}{\|\mathbf{e}_n\|}, & c_{j,n} > 0. \end{cases} \quad (6)$$

As $c_{j,n} < C_{max,n}$, $j \in \bar{\mathcal{A}}_n$, the angles for the coefficients in the inactive set are related to the coefficients in the active set as

$$\arccos \frac{c_{j,n}}{\|\mathbf{e}_n\|} > \arccos \frac{C_{max,n}}{\|\mathbf{e}_n\|}, \quad (7)$$

which yields

$$\left| 90^\circ - \arccos \frac{c_{j,n}}{\|\mathbf{e}_n\|} \right| < 90^\circ - \arccos \frac{C_{max,n}}{\|\mathbf{e}_n\|}, \quad (8)$$

as illustrated in Fig. 2, from where it is possible to conclude that

$$\Delta \theta_n = \max(\theta_n) - \min(\theta_n) = 2 \left(90^\circ - \arccos \frac{C_{max,n}}{\|\mathbf{e}_n\|} \right), \quad (9)$$

where $\theta_n = [\theta_{1,n} \ \cdots \ \theta_{j,n} \ \cdots \ \theta_{J,n}]$.

In [9], a criterion using vector θ_n was proposed to stop the LAR algorithm when $\Delta \theta_n$ reaches a minimum threshold.

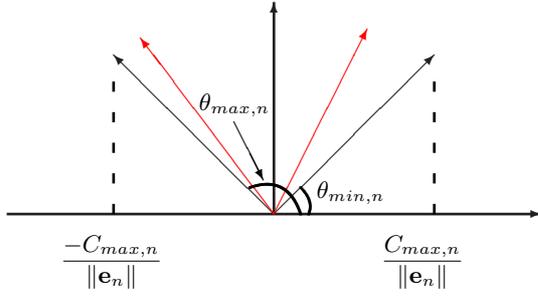


Fig. 2. Active channels ($j \in \mathcal{A}_n$) yield the maximum angle, $\theta_{max,n}$, and the minimum angle, $\theta_{min,n}$. Vectors in red illustrate angles corresponding to channels in the inactive set.

In case of limited amount of data, we propose the improved threshold

$$\Delta\theta_n \leq \mu\sigma_{\theta_1}, \quad (10)$$

where $0 \leq \mu \leq 1$ and σ_{θ_1} is the standard deviation of the angles at the first step. In theory, the value of μ could be set to $\mu > 1$, but practical results show that too few channels are active for $\mu > 1$, resulting in inaccurate coefficient vectors.

When the condition in (10) is satisfied, the LAR algorithm is interrupted and the number of active channels (nonzero coefficients) $n = N$ is determined.

3. THE PROPOSED SCHEME

To allow an efficient use of the RLS algorithm to identify a high order sparse system, we propose the following scheme based on a situational awareness provided by the GLAR algorithm.

Since the LAR algorithm is not adaptive and must be used with a fixed amount of data, we have used data windows of size $L = 5J$; we have observed that with this amount of data the MSE resulted by the LAR algorithm is close to the minimum MSE imposed by the observation noise.

The GLAR algorithm provides the necessary information (number and positions of nonzero coefficients) to the RLS algorithm. With privileged information, the RLS algorithm runs with a reduced complexity at the expense of occasional peaks in computational complexity required by the LAR algorithm when assessing the system.

In order to be robust in face of possible system non-stationarity, a trigger calls the LAR routine. This trigger is implemented via monitoring the estimation error: a variable τ is increased under certain conditions, as shown in Algorithm 1. We use counter τ as a *trigger finger control*, which is increased whenever $|e(k)| > \hat{\sigma}_e$. By $\tau = J$, it is unlikely that J outliers occurred, and τ is increased without checking if $|e(k)| > \hat{\sigma}_e$. When the minimum amount of data $L = 5J$, is gathered, $\tau = 0.5L$ triggers the situational awareness routine.

The proposed scheme is detailed in Algorithm 1. The standard deviation $\hat{\sigma}_e$ of the *a priori* error is estimated continuously with a moving average. Due to the lack of space,

we did not detail algorithms GLAR [1, 9] and RLS [21]. Also refer to [1, 9] for important information on data normalization for the LAR algorithm.

Algorithm 1 – The GLAR-RLS algorithm

```

1:  $L \leftarrow 5J$ 
2: for  $k = 1, 2, 3, \dots$  do
3:   if  $k < L$  then
4:     run the RLS ( $J$  coefficients)
5:   else
6:     if  $k = L$  or  $\tau = 0.5L$  then
7:       run the GLAR, obtain  $N \leq J$ 
8:        $\tau \leftarrow 0$ 
9:     else
10:      run the RLS ( $N$  coefficients)
11:      if ( $\tau < J$  and  $|e(k)| > \hat{\sigma}_e$ ) or ( $\tau \geq J$ ) then
12:         $\tau \leftarrow \tau + 1$ 
13:      end if
14:    end if
15:  end if
16:  update  $\hat{\sigma}_e$ 
17: end for

```

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed algorithm in a sparse system identification, we used the setup depicted in Fig. 3.

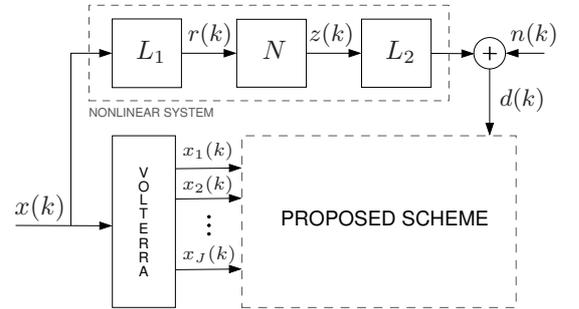


Fig. 3. System setup used to identify the nonlinear system. Block named “PROPOSED SCHEME” may be the RLS or the GLAR algorithm according to Algorithm 1.

The nonlinear system is represented by an LNL model: two linear digital filters (L_1 and L_2) and one memoryless nonlinearity (N). The reference signal $d(k)$ comprises the output of the nonlinear system plus zero mean white noise with variance $\sigma_n^2 = 10^{-6}$. Two scenarios are simulated, with a fixed coefficient vector that changes abruptly in $k_1 = 4,000$ and in $k_2 = 7,000$.

Defining $\mathbf{x}(k) = [x(k) \ x(k-1) \ x(k-2)]^T$ and $\mathbf{z}(k) = [z(k) \ z(k-1) \ z(k-2)]^T$, the LNL model is described as follows. In the first scenario, the system input $x(k)$ is modeled by a white noise with variance $\sigma_x^2 = 1$ and the LNL

model is

$$r(k) = \begin{cases} [0.5 & 1 & 0.5] \mathbf{x}(k), & k < k_1 \\ [0.5 & 0 & 0.5] \mathbf{x}(k), & k_1 \leq k < k_2 \\ [0.5 & 1 & 0.5] \mathbf{x}(k), & k \geq k_2 \end{cases}$$

$$z(k) = r(k) - r^3(k)$$

$$d(k) = \begin{cases} [0.1 & -0.5 & 0.1] \mathbf{z}(k) + n(k), & k < k_1 \\ [0.1 & -0.5 & 0] \mathbf{z}(k) + n(k), & k_1 \leq k < k_2 \\ [0.1 & -0.5 & 0.1] \mathbf{z}(k) + n(k), & k \geq k_2. \end{cases}$$

In the second scenario, the system input $x(k)$ is modeled by a white noise with variance $\sigma_x^2 = 10^{-2}$ and the LNL model is

$$r(k) = [0.5 \ 1 \ 0.5] \mathbf{x}(k)$$

$$z(k) = \begin{cases} \text{tanh}(r(k)), & k < k_1 \\ 0.1r(k) - 0.01r^3(k), & k_1 \leq k < k_2 \\ r(k) - r^3(k), & k \geq k_2 \end{cases}$$

$$d(k) = [0.1 \ -0.5 \ 0.1] \mathbf{z}(k) + n(k).$$

In the first scenario for $k_1 \leq k < k_2$, the simulated system has $N = 12$. In the second scenario for $k < k_1$, N is unknown, since the nonlinearity is given by a hyperbolic tangent. In both scenarios, apart from those periods just pointed out, the simulated system has $N = 27$. Nevertheless, the third-order Volterra filter, with memory equal to four, yields 55 coefficients in the kernel, not counting the DC component.

To compare the results of the GLAR-RLS algorithm, the conventional RLS algorithm with a forgetting factor of $\lambda = 0.99$ is also employed.

The results of a single run are shown in Figs. 4 and 5. Fig. 4 shows the value of N at each iteration for different values of parameter μ ; the real number of nonzero coefficients are indicated in a dashed line. Smaller values of μ yield more coefficients to be estimated. When μ tends to zero, we obtain the response of the RLS algorithm with $N = J$ nonzero coefficients. We can identify in this figure when the GLAR algorithm is called by observing the discontinuities; however, in Fig. 4(b) for $\mu = 0.5$ there is no discontinuity near k_1 , although the GLAR algorithm had been called (the number of estimated coefficients was the same).

As for computational complexity, an overview of this experiment can be seen in Fig. 5 in terms of number of multiplications per time instant k . Every time the GLAR algorithm is run, a peak in the computational complexity appears, as seen in Fig. 5 on the left graphic. However, if the scheme is evaluated dynamically and the computational load is integrated over time, due to the reduced number of coefficients used by the LAR algorithm, the GLAR-RLS scheme turns out to be more economic, always saving energy in the long run. This is observed in Fig. 5 on the right graphic. Results in Fig. 5 are related to the first scenario; the second scenario has a similar result, omitted here due the lack of space.

The average of the MSE over 100 runs is presented in Fig. 6 for the GLAR-RLS algorithm with $\mu = 0.5$ and $\mu = 0.3$ for the first and second scenarios, respectively, as well as

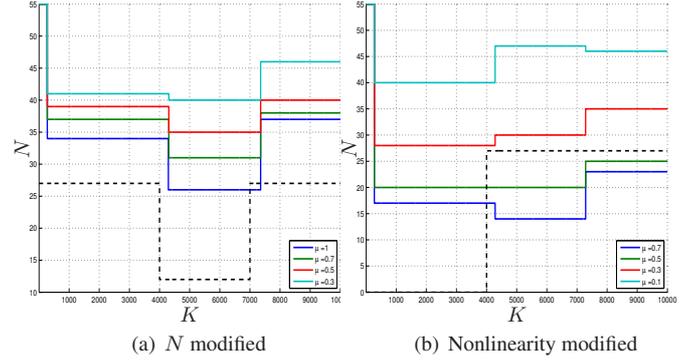


Fig. 4. Coefficients estimated by the proposed algorithm with different values of μ . The real number of nonzero coefficients is given by the dashed line. $k_1 = 4000$ and $k_2 = 7000$.

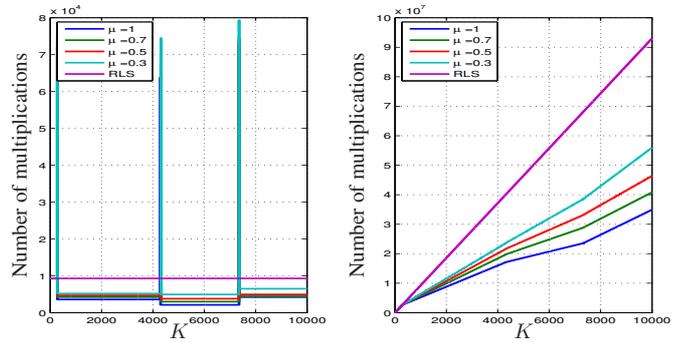


Fig. 5. Computational complexity: instantaneous number of multiplications (left) and summed up over time (right).

for the full (J coefficients) RLS algorithm. Parameter μ is chosen based on the amount of data available (windows of L samples) and in the relation of zero/nonzero coefficients; the resulting N impacts directly on computational complexity. The choice of μ is therefore critical to having amenable performance of the proposed GLAR-RLS scheme in terms of long run efficiency. During the time period that τ is being increased, the MSE result of the GLAR-RLS algorithm is deteriorated when compared to the RLS algorithm, but once the GLAR algorithm is run, the MSE result is enhanced when compared with the full RLS algorithm, since the correct coefficients (or most of them) are set to zero and the recursion in time is over $N < J$ nonzero coefficients. The convergence time of the GLAR-RLS algorithm is shorter than the RLS algorithm after abrupt changes of the plant to be identified.

5. CONCLUSION

In this work, situational awareness based on the LAR algorithm using a geometrical stopping criterion was used to assist the RLS algorithm in identifying sparse systems: after the GLAR algorithm provides privileged information (number and position of nonzero coefficients), the RLS algorithm

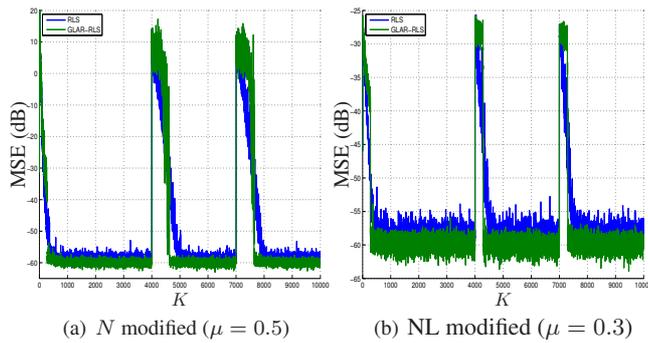


Fig. 6. The resulted MSE, calculated with *a priori* error, for the GLAR-RLS algorithm and for the conventional RLS algorithm (using J coefficients).

runs with reduced computational complexity. The proposed scheme, called the GLAR-RLS algorithm, cascade with a Volterra filter, was used to identify two nonlinear systems. The performance of the GLAR-RLS algorithm was compared with that of the RLS algorithm. In both simulated scenarios, where the system changed abruptly, the GLAR-RLS algorithm presented more accurate estimation, reduced time-averaged computational complexity, and reduced energy consumption over long periods of continuous operation. The proposed scheme also works well in stationary environments, setting the the nonzero coefficients to the RLS algorithm

REFERENCES

- [1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle Regression," in *Annals of Statistics*, Stanford University, 2004, vol. 32, pp. 407 – 409.
- [2] L. Xi, D. Hang, and W. Mingwen, "Two-stage feature selection method for text classification," in *International Conference on Multimedia Information Networking and Security*, Nov. 2009, vol. 1, pp. 234 – 238.
- [3] S. Sathiya Keerthi, "Generalized LARS as an effective feature selection tool for text classification with SVMs," in *22nd International Conference on Machine Learning*, 2005, pp. 417 – 424.
- [4] X. Li, "Finding deterministic solution from underdetermined equation: large-scale performance modeling by least angle regression," in *46th ACM/IEEE Design Automation Conference*, July 2009, pp. 364 – 369.
- [5] C. Xiao, "Two-dimensional sparse principal component analysis for face recognition," in *2nd International Conference on Future Computer and Communication*, May 2010, vol. 2, pp. 561 – 565.
- [6] C. Valdman, M. L. R. de Campos, and J. A. Apolinário Jr., "Nonlinear system identification with LAR," in *International Workshop on Telecommunications*. Instituto Nacional de Telecomunicações, May 2011, pp. 195–200.
- [7] C. Valdman, M. L. R. de Campos, and J. A. Apolinário Jr., "Nonlinear system identification with LAR," *Revista Telecomunicações*, vol. 13, no. 02, pp. 12–21, Dec. 2011.
- [8] D. L. Donoho and Y. Tsaig, "Fast solution of l_1 -norm minimization problems when the solution may be sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.
- [9] C. Valdman, M. L. R. de Campos, and J. A. Apolinário Jr., "A geometrical stopping criterion for the LAR algorithm," in *20th European Signal Processing Conference (EUSIPCO)*, Aug. 2012, pp. 2104–2108.
- [10] B. Dumitrescu, A. Onose, P. Helin, and I. Tăbuș, "Greedy sparse RLS," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2194–2207, May 2012.
- [11] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.
- [12] V. Kekatos and G. B. Giannakis, "Sparse Volterra and polynomial regression models: recoverability and estimation," *IEEE Transactions on Signal Processing*, vol. 9, pp. 5907–5920, 2011.
- [13] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
- [14] G. Schwarz, "Estimating the dimension of a model," in *The Annals of Statistics*, 1978, vol. 6, pp. 461–464.
- [15] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.
- [16] J. A. Khan, S. Van Aelst, and R. H. Zamar, "Robust linear model selection based on least angle regression," *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1289–1299, 2007.
- [17] F. V. Berghen, "LARS Library: Least Angle Regression laSso/Stagewise Library," 2005, Accessed in: 28 Jul. 2010.
- [18] M. Morup, K. H. Madsen, and L. K. Hansen, "Approximate l_0 constrained non-negative matrix and tensor factorization," in *IEEE International Symposium on Circuits and Systems*, May 2008, pp. 1328–1331.
- [19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, Feb. 2009.
- [20] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther, "Forward stagewise regression and the monotone LASSO," *Electronic Journal of Statistics*, vol. 1, pp. 1–29, Mar. 2006.
- [21] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementations*, Springer, 3rd edition, 2008.