A GEOMETRICAL STOPPING CRITERION FOR THE LAR ALGORITHM

Catia Valdman,[1] *Marcello L. R. de Campos,*[1] *and José Antonio Apolinário Jr.*[2]

[1]**Program of Electrical Engineering**
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
email: catia@valdman.com, mcampos@ieee.org

[2]**Department of Electrical Engineering**
Instituto Militar de Engenharia
Rio de Janeiro, Brazil
email: apolin@ime.eb.br

ABSTRACT

In this paper a geometrical stopping criterion for the Least Angle Regression (LAR) algorithm is proposed based on the angles between each coefficient data vector and the residual error. Taking into account the most correlated coefficients one by one, the LAR algorithm can be interrupted to estimate a given number of non-zero coefficients. However, if the number of coefficients is not known *a priori*, defining when to stop the LAR algorithm is an important issue, specially when the number of coefficients is large and the system is sparse. The proposed scheme is validated employing the LAR algorithm with a Volterra filter to identify nonlinear systems of third and fifth orders. Results are compared with three other criteria: Akaike Information, Schwarz's Bayesian Information, and Mallows C_p.

1. INTRODUCTION

The Least Angle Regression (LAR) algorithm was first developed by Efron et al [1]. The LAR algorithm heuristically fits sparse models as a greedy stepwise algorithm [2]. The algorithm constructs a sparse solution to a given problem by iteratively building an approximation, which is even more appropriate for large-scale problems [2]. It was already used successfully in several applications, as, for example, in [3] and [4].

Nonlinear system models are used in many areas, such as communication systems, power amplifiers, loudspeakers with harmonic distortion, and others [5]. The Volterra filter is commonly used to identify nonlinear systems, being a nonlinear filter that may be composed by an infinite number of coefficients [6]. For this reason, standard approaches tend to limit the order of the filter to its second-order, avoiding a large number of coefficients. For example, that was the case when the Volterra filter was used to model nonlinear acoustic echo paths [7] and to identify nonlinear systems [8].

In [9], the authors concluded that by using the LAR algorithm with a Volterra filter, it was possible to identify the most relevant coefficients in nonlinear system modeling, allowing the use of filters with higher orders. Based on the known number of coefficients, the LAR algorithm was interrupted and the correct coefficients were estimated, independently of its Volterra kernel. In this paper, we suppose that the number of non-zero coefficients is not known and the algorithm must be interrupted when the number of coefficients estimated is somehow sufficient; for this task, a geometrical criterion to stop the LAR algorithm is proposed.

Recently, Elad [10] pointed out that the main problem of the LAR algorithm is the tendency to give too many non-zero coefficients. If the algorithm is interrupted when the correct number of non-zero coefficients have been calculated, this problem could be overcome.

Donoho and Tsaig [2] developed a criterion for the Homotopy algorithm, which they called k-step solution property, to be used for under-determined systems. We propose a geometrical criterion to be used in over-determined systems. In order to validate the proposition, we identified nonlinear systems with Volterra filters using the LAR algorithm. We compared performances when the number of nonzero coefficients is known with those when adopting the geometrical criterion based on standard deviation, Mallows C_p, Akaike, and Schwarz's Bayesian criteria.

This paper is organized as follows. Section 2 reviews the LAR algorithm; thereafter, a geometrical criterion based on the angles between each coefficient data vector and the residual error is addressed. The proposed criterion is tested in simulated scenarios, where nonlinear systems were identified using the LAR algorithm and Volterra filters; results are shown in Section 3. Conclusions follow in Section 4.

2. THE GEOMETRICAL STOPPING CRITERION FOR THE LAR ALGORITHM

Before developing the geometrical stopping criterion, the LAR algorithm will be briefly reviewed in order to standardize terminology and nomenclature.

2.1. The LAR Algorithm

After the LAR algorithm was introduced in 2004 [1], several studies were carried out, as can be seen, for example, in [2], [10] and [11].

If the nomenclature used is not well defined, the LAR algorithm can be quite confusing. Therefore, we will first establish some main equations, being as close as possible to the ones used in the field of signal processing.

The input data vector ($J \times 1$) is defined as

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

being k the time index, $k = 1, 2, \cdots, K$, and j the channel index, $j = 1, 2, \cdots, J$. In the present paper, $\mathbf{x}(k)$ is the output of a Volterra filter, as detailed in Section 3. Gathering all input samples from 1 to K, the $K \times J$ input matrix is defined as $\mathbf{X} = [\mathbf{x}(1) \ \ \mathbf{x}(2) \ \ \cdots \ \ \mathbf{x}(K)]^T$, i.e., each row corresponds to the input signal for a given value of k.

The LAR algorithm uses this matrix from another point of view: taking the jth channel, from now on named as coefficient, with all its samples, we may write the input matrix as:

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

where

$$\mathbf{x}_j = [x_j(1) \ \cdots \ x_j(k) \ \cdots \ x_j(K)]^T \tag{3}$$

is the jth coefficient data vector.

All K predictions are also gathered in an output vector, $\mathbf{y}$, as follows:

$$\mathbf{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(K) \end{bmatrix} = \mathbf{Xw}, \tag{4}$$

where $\mathbf{w}$ is the filter coefficient vector.

Being $\mathbf{d}$ the reference vector, $\mathbf{d} = [d(1) \ \ \cdots \ \ d(K)]^T$, and $\mathbf{e}$ the prediction error vector, $\mathbf{e} = \mathbf{d} - \mathbf{y}$, we define the $J \times 1$ correlation vector $\mathbf{c}$, between the input matrix and the prediction error, as follows:

$$\begin{aligned} \mathbf{c} &= \mathbf{X}^T(\mathbf{d}-\mathbf{y}) = [\mathbf{x}(1) \cdots \mathbf{x}(K)] \begin{bmatrix} e(1) \\ \vdots \\ e(K) \end{bmatrix} \\ \begin{bmatrix} c_1 \\ \vdots \\ c_J \end{bmatrix} &= \begin{bmatrix} \sum_{k=1}^K e(k)x_1(k) \\ \vdots \\ \sum_{k=1}^K e(k)x_J(k) \end{bmatrix} \end{aligned} \tag{5}$$

As can be observed from the previous equation, the jth element of $\mathbf{c}$, c_j, corresponds to an estimate of the correlation between the error and the jth input signal. The correlation vector is calculated at every step of the algorithm, identified by $\mathbf{c}_n$, being $n = 1, \cdots, N$ the algorithm iteration, which, for a matter of conciseness, was omitted here.

The set of coefficients inside the model is called the active set, $\boldsymbol{\mathcal{A}}$. The LAR prediction vector ($\mathbf{y}$) proceeds, as the name suggests, in an equiangular direction among the vectors inside the model. The active set starts empty, meaning that all coefficients are equal to zero. The LAR algorithm uses N (up to J) steps to provide a possibly sparse solution with N coefficients computed from the estimated output vector $\mathbf{y}$. The output vector $\mathbf{y}$ is updated according to the following equation ($\mathbf{y}_0 = \mathbf{0}$)

$$\mathbf{y}_n = \mathbf{y}_{n-1} + \gamma_n \mathbf{u}_n, \tag{6}$$

where $n = 1, \cdots, N$ is the algorithm iteration counter, γ_n is the step size, and $\mathbf{u}_n$ is the direction vector. γ_n and $\mathbf{u}_n$ are properly defined in [1], and more detailed in [11].

It is also important to note that all input variables must be normalized before initialization: the coefficient data vector ($\mathbf{x}_j$) must be zero-mean and of unitary length, and the reference vector ($\mathbf{d}$) must be zero-mean.

2.2. The Geometrical Criterion

It is known that at the last possible iteration of the LAR algorithm (when $N = J$) the Least Squares (LS) solution is calculated and the correlation vector is zero [1], such that

$$\mathbf{c} = \mathbf{X}^T\mathbf{e} = 0; \tag{7}$$

therefore,

$$\mathbf{x}_j \perp \mathbf{e} \tag{8}$$

for any j, which is known as orthogonality principle.

But what if we do not estimate all coefficients? What happens to the angle between the coefficient data vector and the error vector before reaching the LS solution? Based on these questions, we rewrite (5) to

$$c_j = \mathbf{x}_j^T\mathbf{e} = \|\mathbf{x}_j\|\|\mathbf{e}\| \cos\theta_j \tag{9}$$

such that

$$\theta_j = \arccos \frac{\mathbf{x}_j^T\mathbf{e}}{\|\mathbf{e}\|}, \tag{10}$$

where θ_j is the angle between the coefficient vector and the error vector. Note that $\|\mathbf{x}_j\|$ was suppressed since it is equal to one, as required by data normalization. The correlation vector is calculated at every step of the algorithm, and so is the angle θ_j. Again, the index n was omitted for a matter of conciseness: a complete formulation of (10) is

$$\theta_{j,n} = \arccos \frac{\mathbf{x}_j^T\mathbf{e}_n}{\|\mathbf{e}_n\|}, \tag{11}$$

which is known to be equal to 90^o at the last step of the algorithm, i.e., when $n = J$, or $\theta_{j,J} = 90^o, j = 1, \cdots, J$.

Since the error vector is also known to decrease as more coefficients are estimated, i.e., included in the active set, it was expected that the fraction in (11) decreases, getting closer to zero at every step, resulting in angles closer to 90^o. It is in practice what happens, as it will be verified in Section 3.

Once we identified that the angles decrease monotonically, a criterion using $\boldsymbol{\theta}_n = [\theta_{1,n} \cdots \theta_{j,n} \cdots \theta_{J,n}]$ was developed and used to *stop the algorithm when*

$$\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}, \tag{12}$$

with $\Delta\boldsymbol{\theta}_n = \max(\boldsymbol{\theta}_n) - \min(\boldsymbol{\theta}_n)$ and $\sigma_{\boldsymbol{\theta}_1}$ is the standard deviation of the angles at the first step. In words, when the range of the angles is equal to or smaller than their initial standard deviation, we assume that all necessary coefficients were estimated and the algorithm is interrupted; each angle needs to be inside a cone limited by the standard deviation of the initial angles.

3. SIMULATION RESULTS

In order to evaluate the performance of the criterion based on the standard deviation of the angles, an experiment with two nonlinear systems, employing the setup depicted in Fig. 1, was carried out.

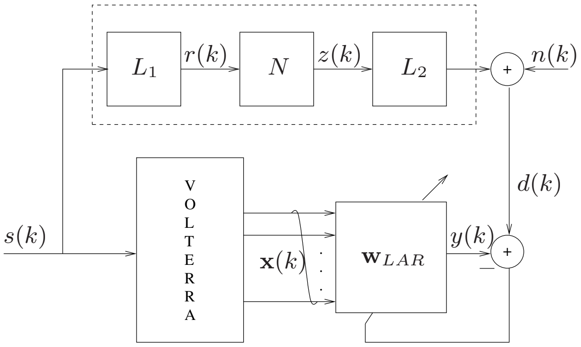


Fig. 1. Nonlinear system identification with Volterra filter and the LAR algorithm.

The nonlinear system was represented by an LNL model, composed by two linear filters with memory (L_1 and L_2) and one nonlinearity (N). Two different scenarios were experimented in order to simulate nonlinear systems of third and fifth orders. The number of Volterra coefficients were 55 and 791, plus the DC component, for the first and second experiments, respectively. In Fig. 1, $n(k)$ is random observation noise.

In order to compare the performance of the standard deviation criterion in (12), we have computed three other criteria [12], [13]: Akaike Information, Schwarz's Bayesian Information, and Mallows C_p (suggested in [1]) criteria, defined as

$$\begin{aligned} AIC &= K\log\left(\frac{\|\mathbf{d}-\mathbf{y}\|^2}{K}\right) + 2n + \frac{2n(n+1)}{K-n+1} \\ BIC &= K\log\left(\frac{\|\mathbf{d}-\mathbf{y}\|^2}{K}\right) + n\log K \\ C_p &= \frac{\|\mathbf{d}-\mathbf{y}\|^2}{\sigma^2_{\mathbf{d}}} - K + 2n \end{aligned}$$

being K the amount of samples and n the number of coefficients estimated. The three criteria work in a similar way: after calculating for all values of n (from 1 to J), the one that yields the smallest AIC/BIC/C_p result is the best number of coefficients to be estimated. This is already a possible disadvantage of these methods: all possible models shall be constructed in order to select one, i.e., the LAR algorithm shall iterate until $n = J$.

3.1. 1st Scenario: Third-Order Volterra + LAR

For the first scenario, the LNL model was constructed as

$$\begin{aligned} L_1: \quad r(k) &= \mathbf{w}_1^T\mathbf{s}(k) \\ N: \quad z(k) &= ar(k) - br^3(k) \\ L_2: \quad d(k) &= \mathbf{w}_2^T\mathbf{z}(k) + n(k) \end{aligned}$$

where

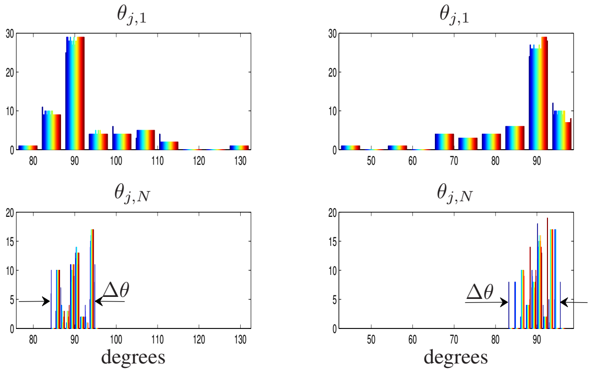
$$\begin{aligned} \mathbf{w}_1 &= [0.5 \quad 1 \quad 0.5]^T \\ \mathbf{s}(k) &= [s(k) \quad s(k-1) \quad s(k-2)]^T \\ \mathbf{w}_2 &= [0.1 \quad -0.5 \quad 0.1]^T \\ \mathbf{z}(k) &= [z(k) \quad z(k-1) \quad z(k-2)]^T \end{aligned}$$

with $\mathbf{w}_1$ and $\mathbf{w}_2$ the first and second filter vectors, respectively, and $\mathbf{s}(k)$, $\mathbf{z}(k)$, and $r(k)$ input signals as depicted in Fig. 1. Although the optimal coefficient vector of the Volterra filter has 55 coefficients, 28 result equal to zero.

In addition, two nonlinear systems were simulated in this scenario, the first, called NL1, had $a = 0.1$ and $b = 0.01$; the second, called NL2, had $a = b = 1$. Their difference is only in the coefficient magnitudes, not in their kernel positions. The main objective here was to evaluate how much the coefficient magnitude influences the results. Both scenarios had a hundred runs averaged to calculate the number of coefficients estimated by each criterion for thirty sets of K, from $K = 100$ to $K = 5,000$.

The histogram of the angles between each coefficient data vector and the error vector is shown in Fig. 2 at the first step and at step N, when $\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}$. Each color means a value of K, from blue to red (there are thirty colors, that is why all seem so close). The angles varied in a range of 50^o for both experiments, being just dislocated when they are compared: from 80^o to 130^o and from 45^o to 95^o, for NL1 and NL2, respectively. At the first step, the angles are spread and at step N they are more concentrated around 90^o; at this point, the prediction error vector is not orthogonal to every coefficient data vector, but the resulting error is already tolerable. When $n = J$, the histogram would be a line at 90^o for both cases, since it corresponds to the LS solution.

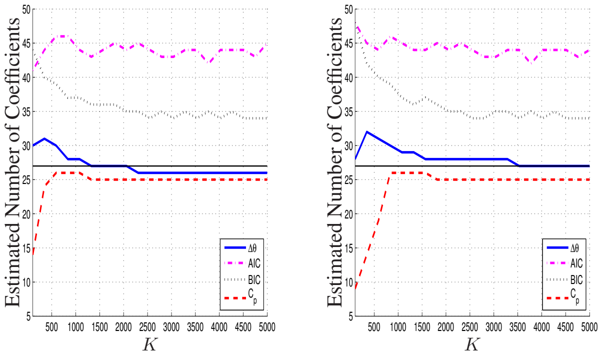
In Fig. 3 we can see the number of coefficients that should be estimated if each of the tested criterion had been chosen to stop the LAR algorithm. As less samples are available, more coefficients are expected to be estimated; however, below approximately $K = 500$, the number of estimated coefficients is increasing in several criteria, meaning that below this value the result is not accurate, the number of samples is not enough



(a) NL1 (b) NL2

Fig. 2. Histogram at the first step and when $\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}$. For each value of K, the initial standard deviation is different, for example, for $K = 100$ we have $\sigma_{\theta_1} = 9.7$ for NL1 and $\sigma_{\theta_1} = 11.2$ for NL2; for $K = 5,000$ we have $\sigma_{\theta_1} = 9.3$ for NL1 and $\sigma_{\theta_1} = 10.5$ for NL2.

to converge. For the LNL system simulated, the number of non-zero coefficients is 27; hence, for both experiments the $\Delta\theta$ criterion is closer to the correct value.



(a) NL1 (b) NL2

Fig. 3. Number of coefficients that should be estimated by each criterion.

3.2. 2nd Scenario: Fifth-Order Volterra + LAR

For the second scenario, the LNL model was constructed as

$$\begin{aligned} L_1: \quad r(k) &= \mathbf{w}_1^T \mathbf{s}(k) \\ N: \quad z(k) &= ar(k) - br^3(k) + cr^5(k) \\ L_2: \quad d(k) &= \mathbf{w}_2^T \mathbf{z}(k) + n(k) \end{aligned}$$

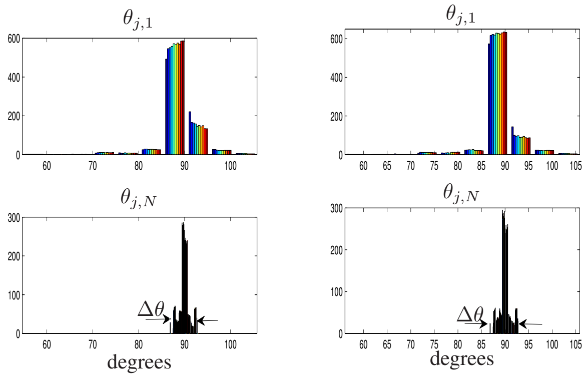
where

$$\begin{aligned} \mathbf{w}_1 &= [0.5 \quad 0.5 \quad 1 \quad 0.5]^T \\ \mathbf{s}(k) &= [s(k) \quad s(k-1) \quad s(k-2) \quad s(k-3)]^T \\ \mathbf{w}_2 &= [0.1 \quad -0.5 \quad 1 \quad -0.5]^T \\ \mathbf{z}(k) &= [z(k) \quad z(k-1) \quad z(k-2) \quad z(k-3)]^T \end{aligned}$$

with $\mathbf{w}_1$ and $\mathbf{w}_2$ the first and second filter vectors, respectively, and $\mathbf{s}(k)$, $\mathbf{z}(k)$, and $r(k)$ input signals as depicted in Fig. 1. Although the optimal coefficient vector of the Volterra filter has 791 coefficients, 580 result equal to zero (the correct number of nonzero coefficients is 211).

Once again, two nonlinear systems were simulated, being NL1 with $a = 0.1, b = 0.01$ and $c = 0.001$ and NL2 with $a = b = c = 1$. Both scenarios had fifty runs averaged to calculate the number of coefficients estimated by each criterion for ten sets of K, from $K = 1,000$ to $K = 10,000$.

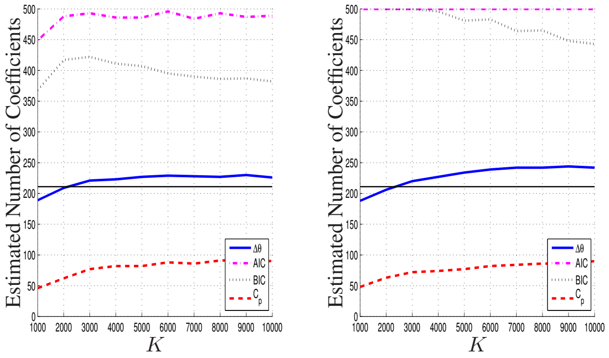
The histogram of the angles are shown in Fig. 4. A behavior similar to the one in the first experiment is seen, angles spread at the first step in a range close to 50^o, whereas at step N, when $\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}$, angles are more concentrated around 90^o.



(a) NL1 (b) NL2

Fig. 4. Histogram at the first step and when $\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}$ for each set of K. For $K = 1,000$ we have $\sigma_{\theta_1} = 4.00$ for NL1 and $\sigma_{\theta_1} = 3.84$ for NL2; for $K = 10,000$ we have $\sigma_{\theta_1} = 3.86$ for NL1 and $\sigma_{\theta_1} = 3.71$ for NL2.

Since estimating all coefficients is very time-consuming and knowing from the LNL model that just 211 coefficients are non-zero, to evaluate AIC/BIC/C_p criteria the LAR algorithm was run until $n = 500$. The number of estimated coefficients by each criteria is shown in Fig. 5. The AIC criterion did not converge for NL2. As in the first experiment, the criterion we propose herein, based on $\Delta\theta$, yields a result which is very close to the correct number of coefficients that should be estimated: approximately 230 for NL1 and 250 for NL2.



(a) NL1 (b) NL2

Fig. 5. Number of coefficients that should be estimated by each criterion.

4. CONCLUSIONS

In this work, a geometrical stopping criterion for the LAR algorithm based on θ_j, angles between the coefficient data vector and the error vector, is proposed to be used in Volterra-based nonlinear system identification. The criterion proposes to stop the LAR algorithm when the range of the angles is equal to or smaller than their initial standard deviation, i.e., $\Delta\boldsymbol{\theta}_n \leq \sigma_{\boldsymbol{\theta}_1}$. This geometrical criterion was compared to Akaike, Schwarz's Bayesian, and Mallows C_p criteria in two different scenarios: identification of nonlinear systems of third and fifth orders. Both nonlinear systems were simulated by LNL models, which were identified using the LAR algorithm in combination with a Volterra filter. In each scenario, two different coefficient magnitude values were tested. One advantage of using the $\Delta\boldsymbol{\theta}$ criterion is that the LAR algorithm does not need to iterate until $n = J$ to evaluate the best value N of non-zero coefficients, as it is needed for the other criteria tested. Also, in all experiments, the geometrical stopping criterion resulted in a number of non-zero coefficients closer to that of the unknown system. In other words, it was able to estimate better the correct number of non-zero coefficients.

Acknowledgements

The authors want to express their gratitude to CAPES, CNPq and INPI-BRAZIL for partial financial support of this work.

5. 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] D.L. Donoho and Y. Tsaig, "Fast solution of l_1 − norm minimization problems when the solution may be sparse," *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 4789–4812, November 2008.

[3] Cuntao Xiao, "Two-dimensional sparse principal component analysis for face recognition," in *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*, May 2010, vol. 2, pp. 561 – 565.

[4] Cuixian Chen, Yaw Chang, K. Ricanek, and Yishi Wang, "Face age estimation using model selection," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, June 2010, pp. 93 – 99.

[5] V. J. Mathew and G. L. Sicuranza, *Polynomial Signal Processing*, John Wiley and Sons, 2001.

[6] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementations*, Springer, 3 edition, 2008.

[7] F. Kuech and W. Kellermann, "Proportionate NLMS algorithm for second-order Volterra filters and its application to nonlinear echo cancellation," in *Conf. Rec. Intl. Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003.

[8] Yoshinobu Kajikawa, "The adaptive Volterra filter: Its present and future," in *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 2000, vol. 83, pp. 51 – 61.

[9] Catia Valdman, Marcello L. R. de Campos, and José A. Apolinário Jr, "Nonlinear system identification with LAR," *Revista Telecomunicações*, vol. 13, no. 02, pp. 12–21, December 2011.

[10] Michael Elad, *Sparse and Redundant Representations*, Springer, 2010.

[11] Jafar A. Khan, Stefan Van Aelst, and Ruben 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.

[12] Dennis J. Beal, "Information criteria methods in SAS for multiple linear regression models," in *Institute for advanced analytics*, 2007.

[13] Kenneth P. Burnham and David R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," in *Sociological methods & research*, November 2004, vol. 33, pp. 261–304.