Shrinkage methods applied to adaptive filters

Marcello L. R. de Campos
Program of Electrical Engineering
Federal University of Rio de Janeiro (UFRJ)
P. O. Box 68504
Rio de Janeiro, Brazil 21941–972
Email: campos@lps.ufrj.br

José A. Apolinário Jr.
Department of Electrical Engineering (SE/3)
Military Institute of Engineering (IME)
Praça General Tibúrcio, 80
Rio de Janeiro, Brazil 22290–270
Email: apolin@ime.eb.br

***Abstract*— This paper analyzes the use of some regression shrinkage methods in adaptive signal processing. Some shrinkage strategies that render interpretable models can be solved as a linearly-constrained least squares problem and render model coefficients which are exactly zero. As a consequence, they produce estimators which may be more economical and have lower variance than those produced by ordinary least squares estimators, at the price of some bias. Economy, in this case, means less computations, consequently less battery consumption and more sustainable systems.**

I. Introduction

In adaptive filtering, the model often used is that of a finite-duration impulse response linear system, where an $M \times 1$ set of signals are fed to the adaptive filter at every iteration. Let $\mathbf{x}(k)$ denote the vector whose M elements, $x_i(k)$, are the input signals to the adaptive filter, and let $y(k)$ be its scalar output signal, all at the time instant kT. M is the length of the filter (therefore its order is $M-1$) and $\mathbf{w}$ is the filter coefficient vector:

$$y(k) = \mathbf{x}^{\mathrm{T}}(k)\mathbf{w}, \tag{1}$$

where $(\cdot)^{\mathrm{T}}$ denotes vector transposition. At every iteration k, an adaptation algorithm must be used to produce (or *estimate*) the elements of $\mathbf{w}(k)$, $w_i(k)$, which are the best for a particular rule, mathematically described by a metric or objective function, with the knowledge available at iteration k. The objective function may use a reference signal, denoted here by $d(k)$, to be pursued by the filter's output; in this case the adaptive filter is said to be supervised. In some applications, usually when constraints are imposed to the filter coefficients, the reference signal is missing and the adaptive filter is called unsupervised. An error $e(k)$ compares the reference and the output signal:

$$e(k) = d(k) - y(k). \tag{2}$$

A. Least Squares Estimation

In the context proposed here, the least squares estimation of $\mathbf{w}$ at time instant kT, which can be traced back to the works of Gauss in the early nineteenth century, relies on past observations of the input and reference signals in order to produce a coefficient vector, $\mathbf{w}(k)$, which satisfies

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left[d(i) - \mathbf{x}^{\mathrm{T}}(i)\mathbf{w}\right]^2. \tag{3}$$

In the previous equation, the minimization is performed over the space $\mathbb{R}^M$, but the extension to the field of complex numbers is trivial.

Let

$$\begin{aligned} &\mathbf{X}(k) = [\mathbf{x}(k)\ \mathbf{x}(k-1)\ \cdots\ \mathbf{x}(1)] \text{ and} \\ &\mathbf{d}(k) = [d(k)\ d(k-1)\ \cdots\ d(1)]^{\mathrm{T}}. \end{aligned} \tag{4}$$

The minimization problem solved by the least squares estimator can be rewritten as

$$\min_{\mathbf{w}} \|\mathbf{d}(k) - \mathbf{X}^{\mathrm{T}}(k)\mathbf{w}\|^2, \tag{5}$$

yielding

$$\mathbf{w}(k) = \left[\mathbf{X}^{\mathrm{T}}(k)\mathbf{X}(k)\right]^{-1}\mathbf{X}(k)\mathbf{d}(k). \tag{6}$$

This solution is the minimum variance unbiased estimate and can be achieved recursively; at each time instant kT, a column is added to matrix $\mathbf{X}(k)$, which means that a rank-one update is made to matrix $\mathbf{R}(k) = \mathbf{X}(k)\mathbf{X}^{\mathrm{T}}(k)$. The *recursive least squares* uses the matrix inversion lemma to obtain an algorithm whose computational complexity is less severe than that of matrix inversion and whose memory requirements are not increasing with k [1]:

$$\begin{aligned} e(k) &= d(k) - \mathbf{x}^{\mathrm{T}}(k)\mathbf{w}(k-1), \\ \mathbf{w}(k) &= \mathbf{w}(k-1) + \frac{e(k)\mathbf{R}(k-1)\mathbf{x}(k)}{1 + \mathbf{x}^{\mathrm{T}}(k)\mathbf{R}(k-1)\mathbf{x}(k)}, \\ \mathbf{R}(k) &= \mathbf{R}(k-1) - \frac{\mathbf{R}(k-1)\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k)\mathbf{R}(k-1)}{1 + \mathbf{x}^{\mathrm{T}}(k)\mathbf{R}(k-1)\mathbf{x}(k)}. \end{aligned} \tag{7}$$

This algorithm needs to be properly initialized with some value for $w_i(0)$, often chosen equal to zero, and a full-rank matrix $\mathbf{R}(0)$. The implication of such initialization is that the objective function becomes

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left[d(i) - \mathbf{x}^{\mathrm{T}}(i)\mathbf{w}\right]^2 + \mathbf{w}^{\mathrm{T}}\mathbf{R}(0)\mathbf{w}. \tag{8}$$

B. A Statistical View of Some Signal Processing Tools

In the field of statistics, the nomenclature is usually different from that encountered in the signal processing literature, but the mathematical tools are often quite similar, if not the same. Input signals are *independent variables*, or *regressors*, and output signals are *dependent variables*, or *regressands*. Input signal correlation is independent variable *collinearity*. However, once we have mapped problems and solutions from one field to the other, successful methods employed by statisticians can be used in signal processing, and vice-versa.

Statisticians and chemometricians have been using to a great extent a variety of tools based on shrinkage, particularly when the volume of observational data is very large, or when collinearity of regressors is high. These include *partial least squares*, *principal components*, *subset selection*, *ridge regression* [2][3], and *least absolute shrinkage* [4].

This paper presents a comparison of some regression shrinkage techniques in signal-processing applications of adaptive filters. The next section presents the basics of regression shrinkage, whereas Section III presents an adaptive filter implementation of the least absolute shrinkage and selection operator (LASSO). Section IV presents simulation results for two different scenarios and Section V presents some preliminary conclusions.

II. Regression Shrinkage

In statistics, it is often desirable to trade bias for variance as a strategy to improve prediction accuracy. Among the strategies available, shrinking the solution coefficient vector away from the least squares estimates yields good results in many situations.

Ridge regressors [5] shrink the solution by adding a small positive quantity to the main diagonal of matrix $\mathbf{X}(k)\mathbf{X}^{\mathrm{T}}(k)$. Although any positive value does the regularization trick, improving matrix conditioning and shrinking the solution, ridge regressors call for a particular value that satisfies

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left[d(i) - \mathbf{x}^{\mathrm{T}}(i)\mathbf{w}\right]^2 \text{ s.t. } \|\mathbf{w}\|_2^2 \leq t, \quad t > 0. \quad (9)$$

A. LASSO

Figures 1 and 2 show contour plots of points having equal error norms for a constrained minimization problem in a two-dimensional case. From Figure 1, it is clear that although the strategy shrinks the solution, it will seldom yield zero coefficients, even for very small values of t. In [4], Tibshirani proposed an alternative shrinkage strategy, the LASSO regressor, for which the minimization problem becomes

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left[d(i) - \mathbf{x}^{\mathrm{T}}(i)\mathbf{w}\right]^2 \text{ s.t. } \sum_{i=1}^{M} |w_i| \leq t. \quad (10)$$

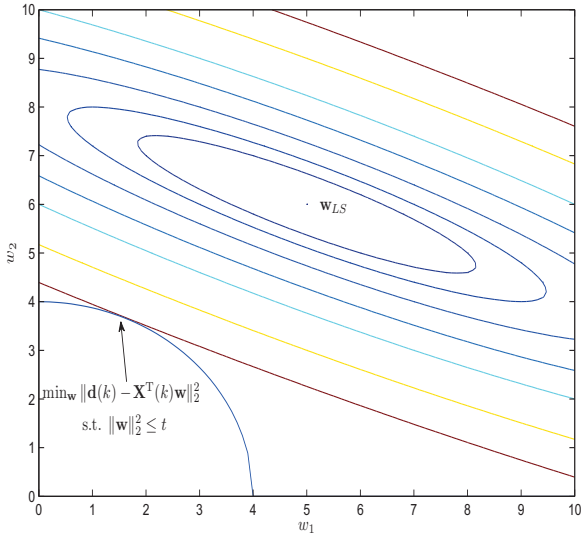

Fig. 1. Least squares and ridge regressor solutions for $t = 16$.

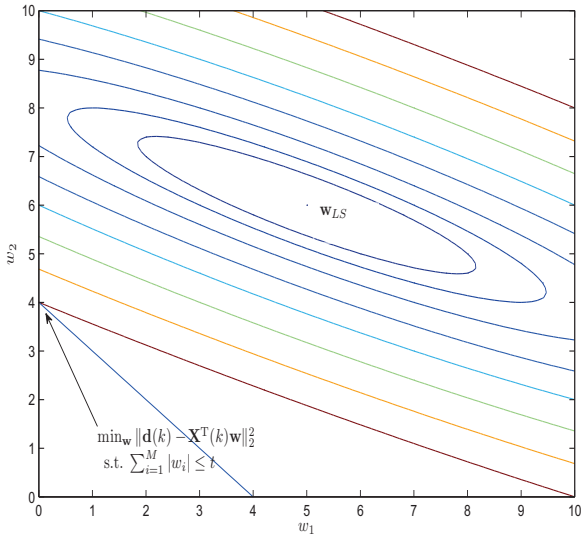

Fig. 2. Least squares and LASSO regressor solutions for $t = 4$.

The LASSO regressor, although quite similar to the ridge regressor, will likely cause some of the coefficients to shrink all the way to zero. Figure 2 illustrates the constrained minimization problem, where one can clearly see that the optimal solution is met when w_1 is zero. Several algorithms have been proposed to solve for the LASSO regressor coefficients (e.g., [6]–[8]). However, a very simple one, albeit not efficient, was proposed by Tibshirani in [4]: Let $\mathbf{s}(\mathbf{w})$ be defined as

$$\mathbf{s}(\mathbf{w}) = \text{sign}(\mathbf{w}). \quad (11)$$

Therefore $\mathbf{s}(\mathbf{w})$ is a member of the set

$$\mathcal{S} = \{\mathbf{s}_j\}, \quad j = 1, \ \ldots, \ 2^M, \quad (12)$$

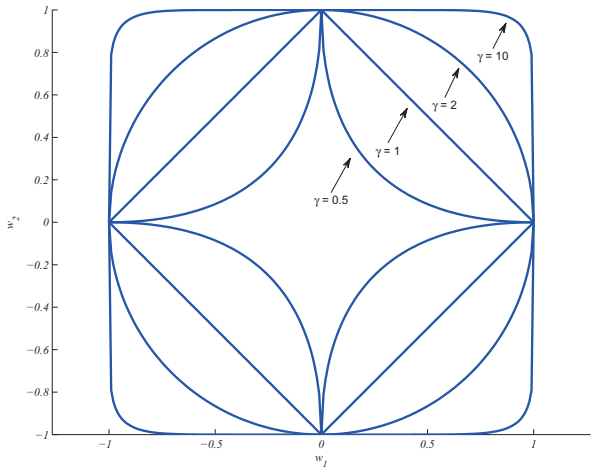

Fig. 3. Boundaries of the constraint sets for different values of γ [2].

whose elements are of the form

$$\mathbf{s}_j = [\pm 1 \ \pm 1 \ \cdots \ \pm 1], \quad j = 1, \ \ldots, \ 2^M. \tag{13}$$

The algorithm is based on the fact that the constraint $\sum_{i=1}^{M} |w_i| \leq t$ is satisfied if

$$\mathbf{s}_j^{\mathrm{T}} \mathbf{w} \leq t, \quad \text{for all } j. \tag{14}$$

The constrained optimization problem becomes a linearly constrained quadratic programming problem whenever the constraint is not satisfied by the least squares estimator.

The constraint equation can be generalized as [2]

$$\sum_{i=1}^{M} |w_i|^{\gamma} \leq t, \quad \gamma > 0. \tag{15}$$

Figure 3 illustrates what happens for the two-dimensional case for different values of γ. When the constraints are active, the solution will likely be more oblique to the coefficient axes, as $\gamma \to \infty$, or aligned with one of the coefficient axes, for $\gamma \leq 1$.

III. Conceptual Adaptive LASSO

As hardware and software design needs to cope with increasingly tighter environmental restrictions, the ability to turning off coefficients automatically may certainly be an advantage worth considering for more economical and greener systems. The fact that the LASSO regressor shrinks coefficients to zero is particularly important for its reduced computational requirements and consequent battery consumption.

Our purpose in this work is to show the potential of applying shrinkage methods in adaptive filtering. At this point, a fully adaptive version of the LASSO algorithm was not our main interest, but the concept and advantages of using it. Therefore, the adaptive LASSO algorithm described herein is subject of further research such that complexity and efficiency are properly tackled.

A simplified adaptive LASSO algorithm is presented here based on Tibshirani's suggested procedure [4]. If $\mathbf{w}_{RLS}(k)$ denotes the least squares estimator, i.e., the coefficient vector that solves Eq. (3), one may obtain the coefficient vector that solves the linearly-constrained least squares problem, $\mathbf{w}_{CRLS}(k)$, as [9][10]

$$\begin{aligned}\mathbf{w}_{CRLS}(k) = \mathbf{w}_{RLS}(k) \\ - \mathbf{R}^{-1}(k)\mathbf{C}(\mathbf{C}^T\mathbf{R}^{-1}(k)\mathbf{C})^{-1}[\mathbf{f} - \mathbf{C}^T\mathbf{w}_{RLS}(k)],\end{aligned} \tag{16}$$

where matrix $\mathbf{C}$ is the constraint matrix and $\mathbf{f}$ is the gain vector. The solution $\mathbf{w}_{CRLS}(k)$ is the vector $\mathbf{w}$ that satisfies

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left[d(i) - \mathbf{x}^{\mathrm{T}}(i)\mathbf{w}\right]^2 \quad \text{s.t. } \mathbf{C}^{\mathrm{T}}\mathbf{w} = \mathbf{f}. \tag{17}$$

The adaptive LASSO algorithm in its conceptual form is presented in Table I where the RLS part is carried out with forgetting factor $\lambda = 1$.

TABLE I
The Conceptual Adaptive LASSO Algorithm.

Initialization:
α (between zero and one)
$\mathbf{1} = [1 \ 1 \ \cdots \ 1]^T$
for each k
{
  % RLS iteration:
  $e_{RLS}(k) = d(k) - \mathbf{w}_{RLS}^T(k-1)\mathbf{x}(k)$
  $\mathbf{k}(k) = \mathbf{R}^{-1}(k-1)\mathbf{x}(k)$
  $\kappa(k) = \frac{\mathbf{k}(k)}{1+\mathbf{x}^H(k)\mathbf{k}(k)}$
  $\mathbf{R}^{-1}(k) = \mathbf{R}^{-1}(k-1) - \kappa(k)\mathbf{k}^H(k)$
  $\mathbf{w}_{RLS}(k) = \mathbf{w}_{RLS}(k-1) + e_{RLS}(k)\kappa(k)$
  % LASSO iteration:
  $t = \alpha \mathrm{sign}(\mathbf{w}_{RLS}^T)\mathbf{w}_{RLS}$
  $e_{LASSO}(k) = d(k) - \mathbf{w}_{LASSO}^T(k-1)\mathbf{x}(k)$
  $\mathbf{w}_{LASSO}(k) = \mathbf{w}_{RLS}(k)$
  $\mathbf{C} = [\ ]$
  while $\mathrm{sign}(\mathbf{w}_{LASSO}^T)\mathbf{w}_{LASSO} > t$
  {
    $\mathbf{C} = [\mathbf{C} \ \ \mathrm{sign}(\mathbf{w}_{LASSO})]$
    $\mathbf{w}_{LASSO}(k) = \mathbf{w}_{RLS}(k) - \mathbf{R}^{-1}(k)\mathbf{C}\left(\mathbf{C}^T\mathbf{R}^{-1}(k)\mathbf{C}\right)^{-1} \times$
    $\left[t\mathbf{1} - \mathbf{C}^T\mathbf{w}_{RLS}(k)\right]$
  }
}

Parameter α in the algorithm of Table I has the effect of controlling the number of coefficients which are equal to zero.

IV. Simulation Results

In order to test the performance of the concepts addressed in the previous section, two experiments were conducted.

In the first experiment, we assumed that the input signal of an adaptive filter of length 50 was formed with signals from 50 different sensors. From these sensors, we assumed

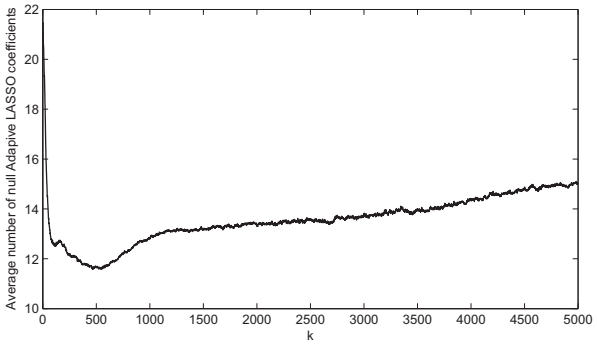

Fig. 4. Number of LASSO zeroed coefficients over time.

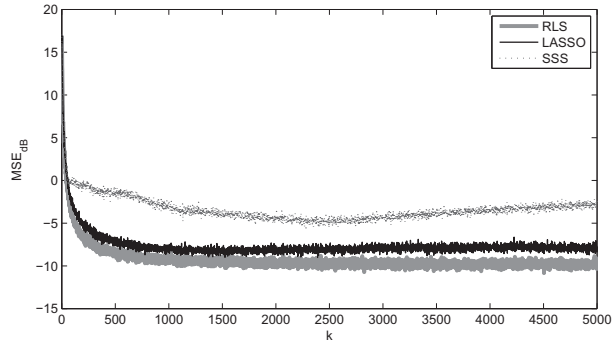

Fig. 5. Learning curves from the faulty coefficients experiment.

that 10% of them, 5 sensors, were defective and were reading only noise.

A second experiment was carried out with the adaptive filter identifying a system having as its impulse response the echo path model 1 used in ITU Recommendation G.168 [11].

For both experiments we implemented the RLS algorithm, the RLS-SSS (Subset Selection) solution, and the adaptive LASSO algorithm. We expected the shrunk versions of the adaptive filter to be able to null some coefficients due to the defective sensors or to their inherent small optimal values. The least squares estimation, needed for all three algorithms, assumed stationary environments and used a forgetting factor equal to one. Further developments of the algorithm may certainly benefit from other weighting options. The subset selection solution is obtained from the ordinary least squares solution, but with a prescribed number of the smallest coefficients (in magnitude) forced to be equal to zero.

A. Experiment with Faulty Sensors

From the 50 sensors signals used to form the input signal vector, we assumed that sensors 10, 20, 22, 39, and 40 were faulty such that only white noise was obtained in these positions. The optimum coefficient vector (unknown plant) was formed from random values uniformly distributed from 0 to 1 and an ensemble of 500 independent runs was carried out.

Figure 4 depicts the number of coefficients zeroed in average by the adaptive LASSO algorithm. Based on this result, the number of zeros set to the SSS algorithm (from the RLS solution) was chosen to be 15.

For this first experiment, a value of t equal to $0.8t_0$ was set. Note that $t_0 = \sum_{i=1}^{M} |w_{RLS_i}|$ changes at each iteration, for $\mathbf{w}_{RLS}(k)$ varies over time. Another possibility, although not very convenient since we do not know $\mathbf{w}_{opt}$ and we would not have a good control of the algorithm, would be a fixed value corresponding, for instance, to $0.8t_{opt}$, with $t_{opt} = \|\mathbf{w}_{opt}\|_1$.

Figure 5 shows the learning curves (MSE in dB) of the three algorithms for this faulty input sensor experiment.

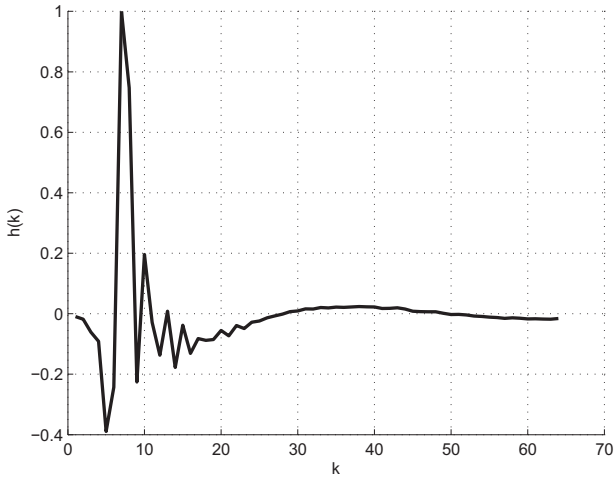

Fig. 6. Impulse response of the ITU-T G. 168 echo path model 1.

Note, as expected, that although having a similar number of null coefficients, the MSE of the adaptive LASSO algorithm exhibits a lower level of MSE. This is due to the fact that it corresponds to the optimal solution constrained to an specific Manhattan norm (leading to a certain number of nulls) while the Subset Selection algorithm corresponds to the RLS solution with the 15 coefficients with smallest magnitudes made equal to zero.

B. ITU-T G.168 Echo Path Identification

A system identification application is carried out with the ITU-T G.168 impulse response model of a long-distance echo path for telephone circuits, shown in Figure 6. This model was chosen for its general availability and for having a long duration with a large tail of values with small magnitude. For this case, the parameter controlling the amount of shrinkage was set to $t = 0.85t_{RLS} = 0.85\|\mathbf{w}_{RLS}\|_1$.

We have once more used the average number of coefficients zeroed by the adaptive LASSO algorithm, as seen in Figure 7, to set the number of zeros in the SSS algorithm, now chosen to be 36.

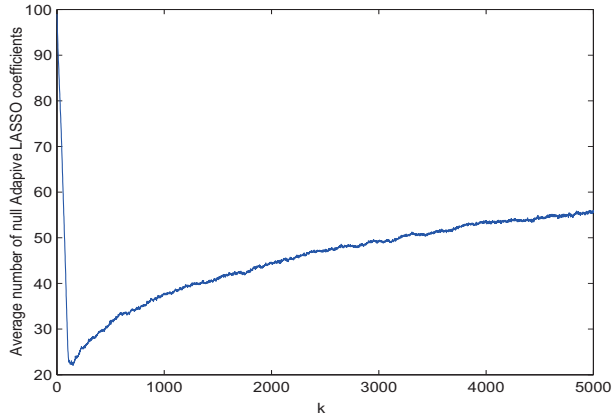

Fig. 7. Number of LASSO zeroed coefficients over time for the case of the ITU-T G.168 echo path identification.

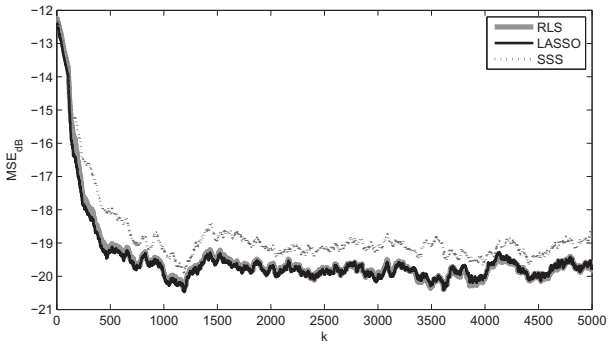

Fig. 8. Learning curves for the ITU-T G.168 echo path identification experiment.

The result of this experiment in terms of MSE is shown in Figure 8 from where we can note that the adaptive LASSO also shows a better behavior than the SSS algorithm.

After convergence, a typical figure with the regressors of the three adaptive algorithms is found in Figure 9. As we can see from this figure, the SSS-RLS algorithm follows exactly the RLS values whenever they are not zero, whereas the adaptive LASSO algorithm presents different values. Nevertheless, in both cases, the samples tend to be zero for the low energy samples of the RLS solution.

V. Conclusion

In this article, we explored the use of shrinkage techniques together with adaptation algorithms in order to make some coefficients equal to zero. We compared the MSE obtained with the LASSO implementation and with a subset selection implementation, which simply replaces coefficients with small absolute value by zero. We tested the concept in two different scenarios. In the first one, some samples of the input signal vector are missing and only additive noise is presented to the adaptive filter, as if some sensors were malfunctioning. In the second scenario,

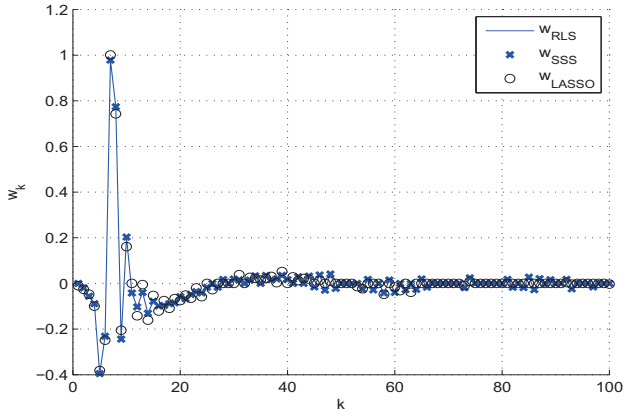

Fig. 9. Adaptive filters regressors for the sparse system identification experiment.

the adaptive filter is to identify a system whose impulse response is very long, but with the energy concentrated in few coefficients. The results for both experiments indicate that the LASSO has a potential to yield shrunk estimates with zero coefficient values and yet acceptable performance.

Acknowledgment

The authors would like to thank CNPq for financing part of this research.

References

[1] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*, 3rd ed. Springer, 2008.

[2] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, May 1993.

[3] Ö. Yeniay and A. Göktas, "A comparison of partial least squares regression with other prediction methods," *Hacettepe Journal of Mathematics and Statistics*, vol. 31, pp. 99–111, 2002.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[5] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, February 1970.

[6] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its dual," *Journal of Computational & Graphical Statistics*, vol. 9, pp. 319–337, June 1999.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, April 2004.

[8] H. Wang, G. Li, and C.-L. Tsai, "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 69, no. 1, pp. 63–78, 2007.

[9] M. L. R. de Campos, S. Werner, and J. A. Apolinário, Jr., "Constrained adaptation algorithms employing Householder transformation," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2187–2195, September 2002.

[10] L. S. Resende, J. M. T. Romano, and M. G. Bellanger, "A fast least-squares algorithm for linearly constrained adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1168–1174, May 1996.

[11] *ITU-T Recommendation G.168: Digital network echo cancellers*. International Telecommunication Union Std., March 2009.