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

Sparse estimation based on a validation criterion

Cristian R. Rojas and Håkan Hjalmarsson

***Abstract*— A sparse estimator with close ties with the LASSO (least absolute shrinkage and selection operator) is analysed. The basic idea of the estimator is to relax the least-squares cost function to what the least-squares method would achieve on validation data and then use this as a constraint in the minimization of the ℓ_1-norm of the parameter vector. In a linear regression framework, exact conditions are established for when the estimator is consistent in probability and when it possesses sparseness. By adding a re-estimation step, where least-squares is used to re-estimate the non-zero elements of the parameter vector, the so called Oracle property can be obtained, i.e. the estimator achieves the asymptotic Cramér-Rao lower bound corresponding to when it is known which regressors are active. The method is shown to perform favourably compared to other methods on a simulation example.**

I. Introduction

One long standing problem in estimation is model selection. In linear regression this amounts to selecting appropriate regressors among a large set of candidate regressors. The brute force approach of comparing all possible subsets using some cross-validation method leads to combinatorial complexity. It is also problematic to analyse the statistical power of this approach.

Many approaches have been suggested to overcome these problems. In Forward Selection regressors are added one by one according to how statistically significant they are [1]. Forward stepwise selection and LARS (Least Angle Regression) [2] are refinements of this idea. Backwards elimination is another approach with a long history. Here regressors are removed one by one. Another class of methods employ all regressors but use thresholding to force insignificant parameters to become zero [3]. Another class of methods that can handle all regressors at once use regularization, i.e. a penalty on the size of the parameter vector is added to the cost function. This approach has close ties with Bayesian estimation. Ridge regression is a classical regularization method where penalty is proportional to the squared 2-norm of the parameter vector. While this method "pulls" parameters towards zero, it does not generate sparse estimates, i.e. even though parameters may be small they are generically non-zero. However, the regularization can be chosen so that sparse estimates are obtained in a one-shot procedure. The LASSO (least absolute shrinkage and selection operator) is one of the early contributions to this field, and has been of tremendous influence[1] [5]. This algorithm performs minimization under a constraint of the ℓ_1-norm of the parameter vector $\theta \in \mathbb{R}^n$. More precisely the criterion is

$$\begin{aligned} &\min_{\theta} \; V_N(\theta) \\ &\text{s.t. } \|\theta\|_1 \leq c \end{aligned} \tag{1}$$

Above $V_N(\theta)$ is the least-squares cost function based on N samples. For linear regression problems the above problem is convex. In fact, one way of viewing (1) is as a convex relaxation of the combinatorial complexity problem of minimizing $V_N(\theta)$ under a constraint of the support of θ.

Using a Lagrange multiplier we see that the LASSO is equivalent to

$$\min_{\theta} \; V_N(\theta) + \lambda \|\theta\|_1 \tag{2}$$

for some $\lambda > 0$. Thus the LASSO can be interpreted as ℓ_1-regularization of the identification criterion. In a Bayesian framework, (2) corresponds to a Laplacian prior.

The reason why (1) gives a sparse estimate is linked to the close relationship between the ℓ_1-norm and the so called ℓ_0-norm, $\|\theta\|_0$ (the number of non-zero entries in θ). With $\|\cdot\|_1$ in (1) replaced by $\|\cdot\|_0$, (1) corresponds to an exhaustive search over all parameters with an upper bound of the number of non-zero parameters.

Integral to many of the approaches is the use of cross-validation or some information criterion, e.g. the Akaike Information Criterion (AIC) or generalized cross-validation (GCV). For example, such methods can be used to determine the constant c in (1). This means solving (1) and then evaluating the performance of the estimate using, e.g., GCV, for different values of c and then picking the best c. While different search strategies for the best c can be deviced, a drawback is that it is necessary to solve (1) multiple times. For large problems this can be restrictive. In this contribution we turn the problem "upside down" and then appeal to AIC to come up with a good way to choose the design parameter (that corresponds to c in (1)). We provide an asymptotic analysis of the proposed estimator. In [6] the finite sample properties of this type of estimator is studied. In [7], a related approach for selecting the regularization parameter for the LASSO is proposed, based on the interpretation of the LASSO as a Maximum a Posteriori estimator, and the use of the Minimum Description Length criterion.

We conclude this introduction by observing that ℓ_1-regularization is closely related to compressive sensing [8].

The outline of the paper is as follows. In Section II the method is introduced together with the assumptions that

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement No. 267381

The authors are with the Automatic Control Laboratory, ACCESS Linnaeus Center, Electrical Engineering, KTH – Royal Institute of Technology, S-100 44 Stockholm, Sweden. Emails: {cristian.rojas|hakan.hjalmarsson}@ee.kth.se

[1] A procedure similar to the LASSO is the *nonnegative garrote* [4]. However, as mentioned in [5], this latter method may perform poorly in overfit or highly correlated settings, while the LASSO and its variants can overcome these issues.

will be used. Section III contains the main results which cover consistency, sparseness and efficiency. The method is illustrated on a numerical example in Section IV, where it is also compared with the LASSO. Conclusions are provided in Section V.

Due to reasons of space, the proofs have been removed. The interested reader is referred to the technical report [9] for the full details of the proofs.

Notation

$X \odot Y$ denotes the Hadamard or element-wise multiplication between two matrices X and Y of the same dimensions. Furthermore, $\|x\|_W^2 := x^T W x$ for $W = W^T > 0$ and $\|x\|_2^2 := x^T x$. $\mathrm{Cond}(A)$ is the condition number of a matrix A in the 2-norm, i.e., $\mathrm{Cond}(A) := \|A\|\|A^{-1}\|$ where $\|A\|$ denotes the maximum singular value of A. Notice that $\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1}) \geq 1$. The vector containing the signs of a vector x is denoted $\mathrm{Sgn}\,[x]$. The pseudo-inverse of a matrix X is denoted $X^\dagger$.

$X_N \xrightarrow{p} X$ denotes convergence in probability [10]. Furthermore, $A_N = O_p(B_N)$ means that, given an $\varepsilon > 0$, there exists a constant $M(\varepsilon) > 0$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{|A_N| \leq M(\varepsilon)|B_N|\} \geq 1 - \varepsilon$. Similarly, $A_N = o_p(B_N)$ means that $A_N/B_N \xrightarrow{p} 0$, and $A_N \asymp_p B_N$ means that, given an $\varepsilon > 0$, there are constants $0 < m(\varepsilon) < M(\varepsilon) < \infty$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{m(\varepsilon) < |A_N/B_N| < M(\varepsilon)\} \geq 1 - \varepsilon$.

In general, all asymptotic statements (of the form $y_N \to y$) are with respect to the number of data samples N tending to infinity.

II. The method and its motivation

A. Data and model

Assumption 2.1 (Data): The data is generated by the linear regression

$$Y_N = \Phi_N \theta^o + E_N \tag{3}$$

where $\theta^o \in \mathbb{R}^n$, $E_N \sim \mathbf{N}(0, \sigma^2 I_N)$ (where $\sigma^2 > 0$), $\Phi_N \in \mathbb{R}^{N \times n}$ and $Y_N \in \mathbb{R}^N$. Furthermore, we assume without loss of generality that $\theta^o = [\theta_1^{oT} \;\; \theta_2^{oT}]^T$, where $\theta_i^o \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_2^o = 0$. We emphasize that this is for notational convenience only; the results below hold regardless of the distribution of zeros in θ^o. The regressor matrix Φ_N is deterministic[2] and satisfies

$$\lim_{N \to \infty} N^{-1} \Phi_N^T \Phi_N =: \Gamma > 0. \tag{4}$$

■

The corresponding model is

$$Y_N = \Phi_N \theta + E_N \tag{5}$$

where $\theta \in \mathbb{R}^n$ is unknown (which also means that $\theta_2^o = 0$ is a priori unknown).

[2]This assumption implies that Φ_N cannot contain autorregressive terms (i.e., past values of the output).

B. The method

Denote the least-squares criterion by

$$V_N(\theta) := \frac{1}{N}(Y_N - \Phi_N \theta)^T (Y_N - \Phi_N \theta). \tag{6}$$

The method we propose for estimating a sparse θ consists of three steps:

i) First compute the ordinary least-squares estimate, $\hat{\theta}_N^{LS}$ say, for the model (5), i.e.

$$\hat{\theta}_N^{LS} = \left(\Phi_N^T \Phi_N\right)^{-1} \Phi_N^T Y_N.$$

ii) Obtain a sparse estimate $\hat{\theta}_N$ solving

$$\begin{aligned} &\min_\theta \; \|\theta\|_1 \\ &\text{s.t. } V_N(\theta) \leq V_N(\hat{\theta}_N^{LS})(1 + \varepsilon_N) \end{aligned} \tag{7}$$

where $\varepsilon_N > 0$. The choice of ε_N will be discussed later.

iii) Finally, re-estimate the non-zero elements of $\hat{\theta}_N$ using ordinary least-squares. More precisely, eliminate the columns of Φ_N in (5) that correspond to zeros in $\hat{\theta}_N$ and then compute the least-squares estimate of a θ of reduced dimension based on the model (5). Thresholding is used to determine which parameters are zero.

When Steps i) and ii) are used, we call this method SPARSEVA (SPARSe Estimation based on VAlidation), the estimate is denoted $\hat{\theta}_N$. When also Step iii) is used, we call the method SPARSEVA-RE, indicating that the non-zero parameters are re-estimated (using least-squares); the corresponding estimate is denoted $\hat{\theta}_N^{RE}$.

For Step ii) we will also consider the following criterion:

$$\begin{aligned} \min_\theta \quad & \|w_N \odot \theta\|_1 \\ \text{s.t.} \quad & (1 + \varepsilon_N) V_N(\hat{\theta}_N^{LS}) \geq V_N(\theta), \end{aligned} \tag{8}$$

where $w_N \in \mathbb{R}_+^n$ is given by $w_{Ni} := 1/|\hat{\theta}_{Ni}^{LS}|^\gamma$ $(i = 1, \ldots, n)$, where $\gamma > 0$ is arbitrary. We denote the method obtained from Step i) and (8) by A-SPARSEVA (Adaptive SPARSEVA) and the corresponding estimate by $\hat{\theta}_N^A$; the method when all three steps is in this case denoted A-SPARSEVA-RE and the corresponding estimate by $\hat{\theta}_N^{A-RE}$. This adaptive version is inspired by the adaptive LASSO [11].

We notice that both (7) and (8) are convex for linear regression problems.

C. Discussion of the method

The idea behind SPARSEVA is based on Akaike's Information Criterion AIC. Let $V_N^{val}(\theta)$ denote the same least-squares cost function as $V_N(\theta)$ but using a fresh validation data set (with the same Φ_N but with a different realization of the noise E_N). Then, for linear regression problems, c.f. [12], it is easily shown that

$$\mathrm{E}_{val}\left[\mathrm{E}_{est}\left[V_N^{val}(\hat{\theta}_N^{LS})\right]\right] = \left(1 + \frac{2n}{N}\right) \mathrm{E}_{est}\left[V_N(\hat{\theta}_N^{LS})\right] \tag{9}$$

where $\mathrm{E}_{est}[\cdot]$ ($\mathrm{E}_{val}[\cdot]$) denotes expectation with respect to the noise in estimation (validation) data set.

The relation (9) suggests that a way to perform model selection without using a validation data set is to minimize

$$\left(1+\frac{2n}{N}\right)V_N(\hat{\theta}_N^{LS})$$

with respect to n, the number of estimated parameters. This is Akaike's AIC criterion for model selection.

In view of this, with the choice $\varepsilon_N = 2n/N$, (7) can be seen as a way to estimate a sparse (due to the ℓ_1-norm) model such that its performance is similar to that of the least-squares estimate on validation data. Thus, unlike the LASSO, there is a natural choice of the "regularization" parameter ε_N for SPARSEVA. This is the motivation for introducing (7).

It should be noted that the criterion

$$\begin{aligned} &\min_{\theta} \ \|\theta\|_1 \\ &\text{s.t. } \ V_N(\theta) \leq \varepsilon \end{aligned} \tag{10}$$

has been used before for signal recovery in a compressive sensing context [13, 14], i.e. when the number of observations N is less than the number of estimated parameters n. Our contributions lie in the suggestion to use ε according to (7), in particular with ε_N chosen according to the AIC-rule $\varepsilon_N = 2n/N$, an asymptotic (in N) analysis, and the adaptive version (8) inspired by [11].

III. Main results

In this section we present the main technical results.

A. Consistency

In regards to consistency we have the following result.

Theorem 3.1 (Consistency of (A-)SPARSEVA): Under Assumption 2.1, and $\theta^o \neq 0$, SPARSEVA and A-SPARSEVA are consistent in probability (i.e.[3], $\hat{\theta}_N^{(A)} \xrightarrow{p} \theta^o$) if and only if $\varepsilon_N \to 0$. In particular, $\|\hat{\theta}_N^{(A)} - \theta^o\|_2 = O_p(N^{-1/2} + \sqrt{\varepsilon_N})$.

Proof: See [9]. ■

Corollary 3.1 (Exact order of consistency): Subject to the assumptions of Theorem 3.1, if $\varepsilon_N \to 0$ but $N\varepsilon_N \to \infty$, then $\|\hat{\theta}_N^{(A)} - \theta^o\|_2 \asymp_p \sqrt{\varepsilon_N}$.

Proof: See [9]. ■

B. Sparseness

Since $V_N(\theta)$ is quadratic, the constraint in (7) is an ellipsoid. The solution to (7) will be on the boundary of the smallest ℓ_1-ball that intersects this ellipsoid, see Figure 1.a. When the ellipsoid has the shape as in Figure 1.a, then, as can be seen, the solution will be sparse. However, with a more tilted ellipsoid as in Figure 1.b, the solution will not be sparse. The shape of the ellipsoid is determined by the regressor matrix Φ_N.

Various measures to ensure sparsity have been suggested, e.g. [15, 16]. The adaptive SPARSEVA (8) is inspired by [11]. We now establish the exact conditions on ε_N for the adaptive SPARSEVA to generate sparse estimates.

Theorem 3.2 (Sparseness of the adaptive SPARSEVA): Under Assumption 2.1, and in addition $\varepsilon_N \to 0$ and $\theta^o \neq 0$. Then, A-SPARSEVA (8) satisfies the sparseness property (i.e., $\hat{\theta}_N^A = [(\hat{\theta}_N^{A1})^T \ (\hat{\theta}_N^{A2})^T]^T$, with $\hat{\theta}_N^{Ai} \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, where $P\{\hat{\theta}_N^{A2} = 0\} \to 1$) if $N\varepsilon_N \to \infty$. If $N\varepsilon_N \to \infty$ does not hold, A-SPARSEVA does not have the sparseness property.

Proof: See [9]. ■

Remark 3.1: It can be shown that when the regressors are orthonormal, i.e. $N^{-1}\Phi_N^T\Phi_N = I$, then Theorem 3.2 holds also for SPARSEVA. ■

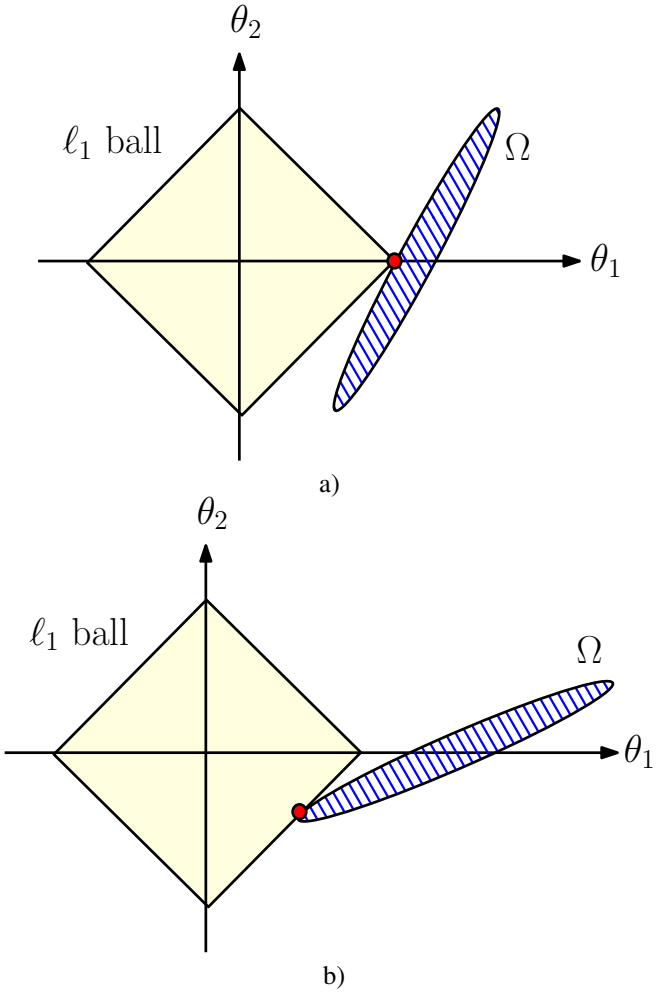

Fig. 1. The geometry of (7). In a) a sparse solution ($\theta_2 = 0$) is obtained but not in b).

C. Adaptive SPARSEVA and the Oracle property

From the preceeding results, the adaptive SPARSEVA possesses the sparseness property if and only if ε_N is chosen such that $\varepsilon_N \to 0$ and $N\varepsilon_N \to \infty$. On the other hand, by Corollary 3.1, such a choice of ε_N gives rise to a non efficient estimator (since the order of convergence of $\hat{\theta}_N^A$ to θ^o would be $\sqrt{\varepsilon_N}$, strictly larger than $N^{-1/2}$). One way to overcome this efficiency-sparseness tradeoff is to add Step iii) (see Section II) so that the non-zero parameters are re-estimated using least-squares. Our next result shows that the estimator obtained from the third step of the adaptive SPARSEVA is asymptotically normal and efficient.

Theorem 3.3 (The Oracle property): Consider the assumptions in Theorem 3.2 and that $N\varepsilon_N \to \infty$. Then

$$\sqrt{N}(\hat{\theta}_N^{A-RE} - \theta^o) \in AsN(0, M^{\dagger})$$

[3]The notation $\hat{\theta}_N^{(A)}$ refers either to $\hat{\theta}_N$ or $\hat{\theta}_N^A$, depending on the context.

where M is the information matrix when it is known which elements of θ^o are zero.

Proof: See [9]. ■

Remark 3.2: We remark that it is clear from the proof of Theorem 3.3 that such result holds if we replace the use of $\hat{\theta}_N^A$ as an estimator of the location of the non-zero components of θ^o by any other consistent estimator of such components. For example, Remark 3.1 implies that Theorem 3.3 holds for SPARSEVA-RE when the regressors are orthonormal. ■

IV. NUMERICAL EXAMPLES

A. Example I

In this section we will illustrate SPARSEVA and compare it with other methods using Example 4.1 in [15]. In this example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

The noise is zero mean, unit variance white Gaussian noise. The regressors are mutually independent, with each regressor being a realization of the output of a first order filter with pole in 0.5 subject to a zero mean Gaussian white input. Each regressor is normalized to have variance 1.

SPARSEVA is compared to the following methods: LS-ORACLE is the least-squares estimate of the three non-zero parameters. This is the ideal estimator and from the Cramér-Rao lower bound no other unbiased estimator can perform better. LASSO-GCV is the LASSO where the regularization parameter λ in (2) is chosen according to generalized cross-validation [5], i.e. the λ that minimizes

$$V_N(\hat{\theta}_N)/(1-p(\lambda)/N)^2$$

is chosen. Here $p(\lambda)$ is the number of effective parameters defined as

$$p(\lambda) = \mathrm{Tr}\left\{\Phi_N\left(\Phi_N^T\Phi_N + \lambda W^\dagger\right)^{-1}\Phi_N^T\right\}$$
$$W = \mathrm{Diag}(|\hat{\theta}_{Ni}|).$$

Four variants of SPARSEVA are included: SPARSEVA-AIC/BIC where the constraint ε_N is chosen as AIC ($\varepsilon_N = 2n/N$) and BIC ($\varepsilon_N = n\log N/N$). A-SPARSEVA-AIC/BIC are the two corresponding adaptive versions. Notice that the BIC choice for ε_N satisfies the condition for sparseness (see Theorem 3.2).

Figure 2 shows the MSE (Mean-Squared Error) of the parameter estimate as a function of the sample size for 100 Monte-Carlo simulations. Re-estimation is used for the SPARSEVA-methods. The threshold for determining which parameters are zero and non-zero, respectively, was (somewhat arbitrarily) set to 10^{-5}. Also re-estimation was tried for LASSO-GCV but was found to perform worse than no re-estimation and has therefore not been included. It can be seen that above $N = 70$, the adaptive SPARSEVA with BIC constraint achieves the Oracle property and performs as well as LS-ORACLE. Figure 3 shows the average number of correctly estimated zero parameters, and we see that this estimator has the best ability to determine where the zero elements are located. However, from Figure 2 it can be seen that for small sample sizes the performance of this estimator is worse than almost all other estimators. From Figure 4,

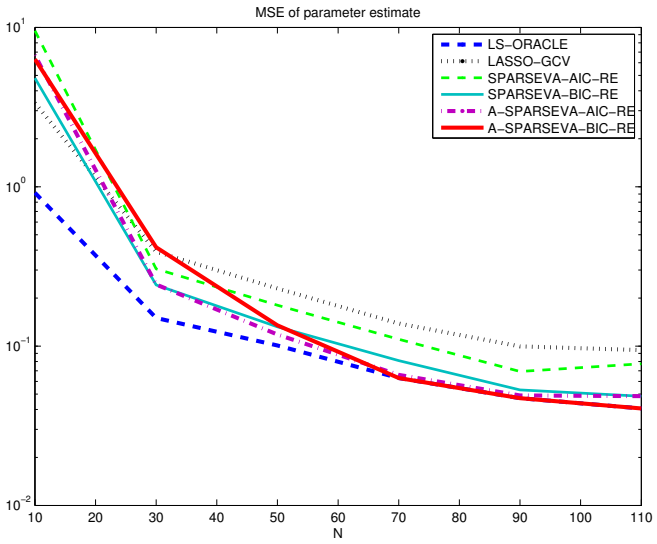

Fig. 2. Example I: MSE as a function of the sample size N.

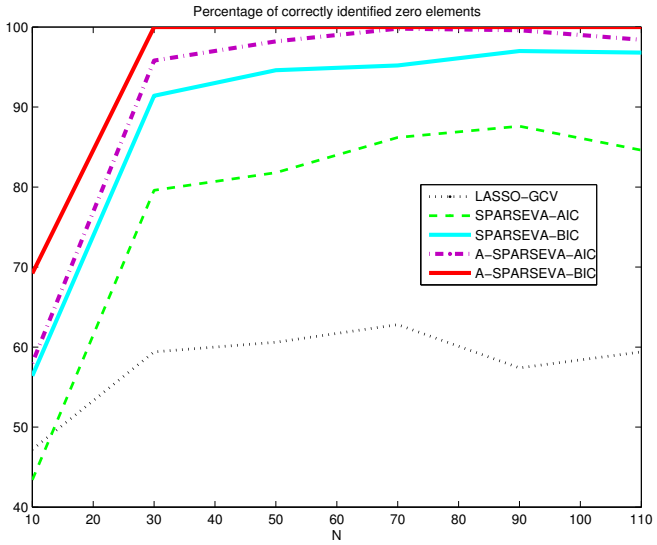

Fig. 3. Example I: Percentage of correctly identified zero elements as a function of the sample size N.

which shows the average number of correctly estimated non-zero parameters, it is clear that this is due to that this estimator has problems to identify which elements of θ^o are non-zero for small sample sizes.

B. Example II

In this section we will consider a more extreme case than in Section IV-A. We will consider the following dynamical finite impulse response model

$$y_t = \sum_{k=0}^{n-1}\theta_k u_{t-k} + e_t$$

where $n = 20$. The true system can be described by this model with $\theta_k = \delta_k$ (δ_k is Kronecker's delta function) and $\{e_t\}$ being Gaussian white noise with variance 1. The input

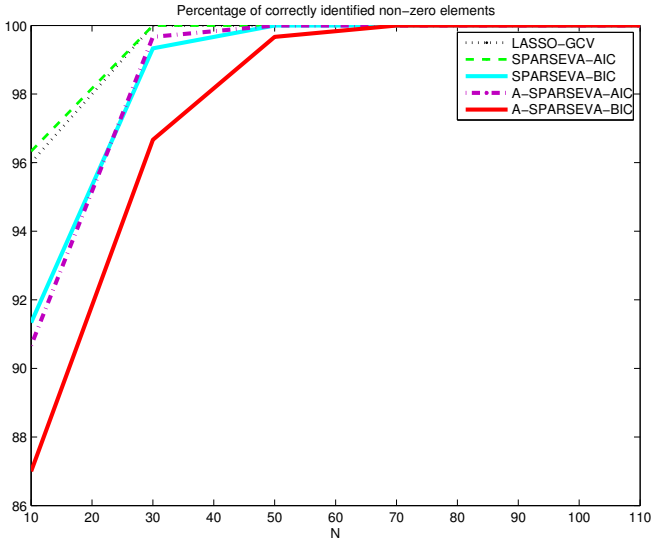

Fig. 4. Example I: Percentage of correctly identified non-zero elements as a function of the sample size N.

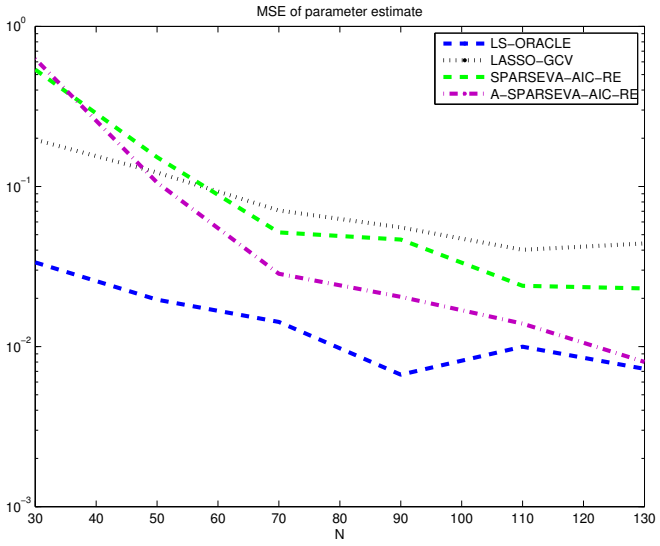

Fig. 5. Example II: MSE as a function of the sample size N.

$\{u_t\}$ is given by

$$u_t = 0.5u_{t-1} + w_t$$

where $\{w_t\}$ is zero mean white noise. The variance of w_t is such that the input has unit variance.

We are thus in the situation where all parameters are zero except the first one. Figure 5 shows the MSE (Mean-Squared Error) of the parameter estimate as a function of the sample size for 50 Monte-Carlo simulations. The included methods are: least-squares oracle, Lasso with GCV, SPARSEVA-RE-AIC and A-SPARSEVA-RE-AIC. For the small sample sizes used the BIC versions of SPARSEVA perform very poorly in this example. The non-zero element is not detected in any of the realizations.

The threshold for determining which parameters are zero and non-zero, respectively, was (somewhat arbitrarily) set to 10^{-5} as in the previous example.

Figure 6 shows the average number of correctly estimated zero parameters. Figure 7 shows the average number of correctly estimated non-zero parameters.

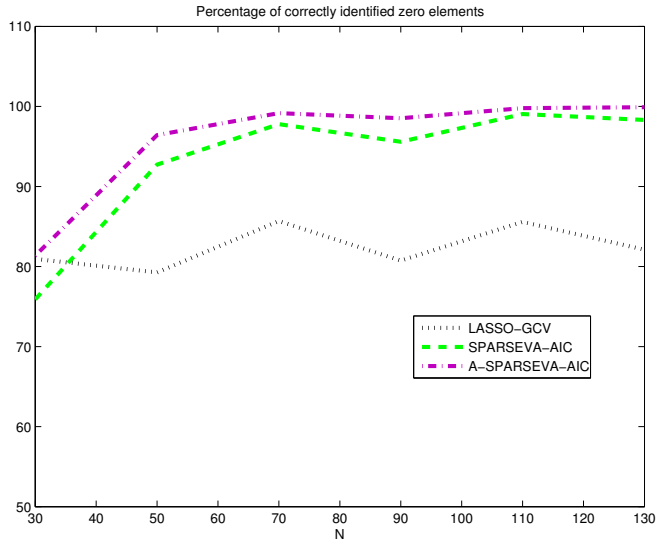

Fig. 6. Example II: Percentage of correctly identified zero elements as a function of the sample size N.

V. CONCLUSIONS

In the numerical examples the adaptive version of the method performs most favourably. On these examples, the "AIC" choice $\varepsilon_N = (1 + 2n/N)V_N(\hat{\theta}_N^{LS})$ seems to give a good balance between sparsity and model fit. Thus, this method has the potential to provide a good estimate in one shot. This is an attractive property for large scale problems where the possiblity to optimize the design parameter may be restricted.

When the focus is on sparseness, the "BIC" choice $\varepsilon_N = (1 + 2\log(N)/N)V_N(\hat{\theta}_N^{LS})$ ensures this property.

ACKNOWLEDGEMENT

The authors would like to thank R. Tóth for an interesting exchange of ideas.

REFERENCES

[1] S. Weisberg, *Applied Linear Regression*. New York: Wiley, 1980.

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

[3] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[4] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37(4), pp. 373–384, 1995.

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

[6] R. Toth, B. S. Sanandaji, K. Poolla, and T. L. Vincent, "Compressive system identification in the linear

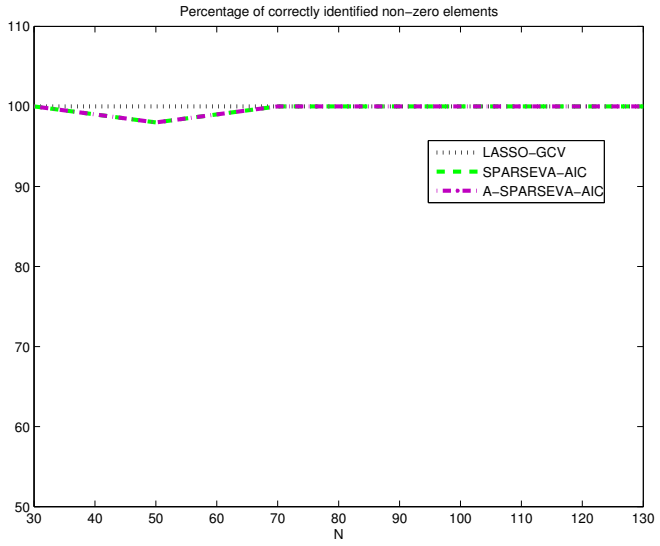

Fig. 7. Example II: Percentage of correctly identified non-zero elements as a function of the sample size N.

time-invariant framework," in *Proceedings 40th IEEE Conference on Decision and Control*, Orlando, Florida, USA, December 2011, submitted.

[7] A. Panahi and M. Viberg, "Maximum a posteriori based regularization parameter selection," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, Czech Republic, May 22-27, pp. 2452–2455.

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

[9] C. R. Rojas and H. Hjalmarsson, "Sparse estimation based on a validation criterion," *http://www.ee.kth.se/˜crro/sparseva.pdf*, 2011, technical Report.

[10] E. L. Lehmann, *Elements of Large-Sample Theory*. Springer, 1999.

[11] H. Zou, "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101(476), pp. 1418–1429, 2006.

[12] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[13] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, pp. 1207–1223, 2006.

[14] A. Gurbuz, J. McClellan, and W. Schott Jr, "Compressive sensing for subsurface imaging using ground penetrating radar," *Signal Processing*, vol. 89, pp. 1959–1972, 2009.

[15] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[16] H. Wang and C. Leng, "Unified LASSO estimation by least squares approximation," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 1039–1048, 2007.