Sparse estimation of polynomial and rational dynamical models

Cristian R. Rojas, *Member, IEEE*, Roland Tóth, *Member, IEEE*, and Håkan Hjalmarsson *Fellow, IEEE*

***Abstract*—In many practical situations, it is highly desirable to estimate an accurate mathematical model of a real system using as few parameters as possible. At the same time, the need for an accurate description of the system behavior without knowing its complete dynamical structure often leads to model parameterizations describing a rich set of possible hypotheses; an unavoidable choice, which suggests sparsity of the desired parameter estimate. An elegant way to impose this expectation of sparsity is to estimate the parameters by penalizing the criterion with the ℓ_0 "norm" of the parameters. Due to the non-convex nature of the ℓ_0-norm, this penalization is often implemented as solving an optimization program based on a convex relaxation (*e.g.*, ℓ_1/LASSO, nuclear norm, ...). Two difficulties arise when trying to apply these methods: (1) the need to use cross-validation or some related technique for choosing the values of regularization parameters associated with the ℓ_1 penalty; and (2) the requirement that the (unpenalized) cost function must be convex. To address the first issue, we propose a new technique for sparse linear regression called SPARSEVA, with close ties with the LASSO (least absolute shrinkage and selection operator), which provides an automatic tuning of the amount of regularization. The second difficulty, which imposes a severe constraint on the types of model structures or estimation methods on which the ℓ_1 relaxation can be applied, is addressed by combining SPARSEVA and the Steiglitz-McBride method. To demonstrate the advantages of the proposed approach, a solid theoretical analysis and an extensive simulation study are provided.**

I. Introduction

System identification is a discipline that deals with the problems of estimating models of dynamic systems from input-output data. Even though its birth is dated back in the era of classical automatic control during the 60's and 70's, by now it has become a mature field with many successful applications in areas such as economics, mechatronics, ecology, biology, communications and transportation [1, 2, 3, 4]. It also has a close connection with allied fields such as statistics, econometrics, machine learning and chemometrics [5].

For a system identification procedure to be successful, two main ingredients are needed: data containing measured information about the dynamics of the system, and prior knowledge. Data is provided by an identification experiment, while the prior knowledge has to be supplied (directly or implicitly) by the user, in the form of assumptions or prejudices. One of the most important prejudices is the selected model structure and the corresponding model set within which the identification method should find an estimate of the plant. Such a selection is rather complicated as it is outmost desired to estimate an accurate model of the real system using as few parameters as possible. While accuracy is clearly related to the performance of the application on which the model will be used, the desire for a minimal parametrization is based on the parsimony principle (Occam's razor) and the utilization complexity in terms of control synthesis, prediction, etc. Since an optimal choice in this question is rarely known a priori, a user in identification typically proposes a model structure capable of explaining a rich set of possible dynamics, and lets the data decide which sub-structure is appropriate to use. This is commonly achieved by employing model structure selection tools (such as AIC, BIC/MDL, cross-validation, etc.). These tools can be seen as imposing a sparsity pattern on the parameters, because they determine a model sub-structure (where the estimated model should be found), by forcing some of the parameters of the overall model structure to be exactly equal to zero. Therefore, model structure selection can be interpreted as the process of imposing a *sparsity prejudice*.

Many techniques for sparse estimation have been successfully used for model structure selection in linear regression settings. For example, in *Forward Selection* regressors are added one by one according to how statistically significant they are [6]. *Forward Stage-wise Selection* and *Least Angle Regression* (LARS) [7] are refinements of this idea. *Backward Elimination* is another approach with a long history. Here regressors are removed one by one based on the same concept of statistical significance. Another class of methods employ all regressors, but use thresholding to force insignificant parameters to become zero [8]. In [9], a Bayesian approach to sparse estimation is developed. Yet 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. The *Least Absolute Shrinkage and Selection Operator* (LASSO) [10] and the *Non-Negative Garrote* (NNG) [11], are early approaches based on the idea of using regularization to enforce sparsity. The LASSO, for example, is based on the minimization of a least-squares cost function plus the ℓ_1 norm of the parameter vector (which is known to enforce sparsity). More precisely the LASSO criterion is

$$\min_{\theta\in\mathbb{R}^{\cdot}} \quad V(\theta, \mathcal{D}_N), \tag{1a}$$

$$\text{s.t.} \quad \|\theta\|_1 \leq \varepsilon, \tag{1b}$$

where V is the least-squares cost function based on a data set $\mathcal{D}_N$ with N samples. For linear regression setups, 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 on the size of the support of θ.

Integral to many of the approaches is the use of cross-validation or some information criterion, *e.g.*, the *Akaike*

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.

C. R. Rojas and H. Hjalmarsson are with the Automatic Control Lab and ACCESS Linnaeus Center, Electrical Engineering, KTH-Royal Institute of Technology, S-100 44 Stockholm, Sweden. R. Tóth is with the Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. Emails: cristian.rojas@ee.kth.se, R.Toth@tue.nl,hakan.hjalmarsson@ee.kth.se.

Information Criterion (AIC) or *Generalized Cross-Validation* (GCV). For example, such methods can be used to determine the constant ε in (1). This means solving (1) and then evaluating the performance of the estimate using, *e.g.*, GCV, for different values of ε and then picking the best ε. While different search strategies for the best ε can be derived, a drawback is that it is necessary to solve (1) multiple times [12, 13]. For large problems, this can be restrictive. As a first contribution, we turn the problem "upside down," starting with a linear regression structure, and then appeal to AIC to come up with an "efficient" way to choose the design parameter (which corresponds to ε in (1b)). We provide an asymptotic analysis of the proposed estimator, called SPARSEVA, which was originally proposed in [14].

An additional complication with the LASSO and most sparse estimation methods is that they can only be applied to model structures of a linear regression type (*i.e.*, where the cost function to be minimized by the estimator is quadratic in the parameters). Some extensions, however, have been conceived for estimators based on the minimization of a convex loss function [15, Chapter 8]. This class of estimators can be easily implemented by using convex optimization tools. For estimators arising from a non-convex loss function, it is much more difficult to impose sparsity, because their implementation can suffer from multiple local minima [15, Chapter 9].

Confinement to estimators with a convex loss function (identification criterion) is very restrictive. This is because, in prediction error minimization, many *Linear Time-Invariant* (LTI) model structures (such as ARMAX, Output-Error, and Box-Jenkins [2]) give rise to a non-convex loss function of the prediction. Even model structures for which this prediction error function is known to have a single global minimum (*e.g.*, ARMA structures [2]) may end up having multiple local optima if an ℓ_1 regulation term is added to it to impose sparsity.

In this paper, our second contribution is to extend the use of convex relaxation techniques for sparsity to general LTI rational *Output Error* (OE) type of model structures estimated using *Prediction Error Methods* (PEM), where we allow the noise to be colored. To this end, we first combine SPARSEVA, and the *Steiglitz-McBride method*, which is a technique for the estimation of OE models. Since the Steiglitz-McBride approach reduces the estimation problem of OE models to solving a sequence of least-squares estimation problems, which are convex optimization programs, we can apply a LASSO penalty to this sequence. This allows to impose sparsity in the resulting plant model, in case the output noise is white. We also extend this approach to general colored noise situations by using a pre-filtering approach with a high-order ARX, which is a recently proposed extension of the Steiglitz-McBride method [16].

The paper is organized as follows. The notation used in the sequel is described in Section II. Section III introduces the problem formulation. A description of the techniques proposed is given in Section IV, where we present the SPARSEVA approach (for linear regression), revisit the classical Steiglitz-McBride method, and describe a technique for the sparse estimation of rational plant models, called OE-SPARSEVA, based on the combination of the first two methods. In Section V, we establish the theoretical asymptotic properties of SPARSEVA and its variants. Section VI presents several simulation examples that show the properties of our proposed methods. Finally, the paper is concluded in Section VII. For the reader's convenience, most proofs have been collected in the appendices.

II. Notation

$X \odot Y$ denotes the Hadamard or element-wise multiplication between two matrices X and Y of the same dimensions. Furthermore, $\|x\|_W^2 := x^\top W x$ for $W = W^\top > 0$, $\|x\|_2^2 := x^\top x$ and $\|x\|_1 = \sum_{i=1}^n |x_i|$ with $x = [x_1 \ \dots \ x_n]^\top$. $\mathrm{Cond}(A)$ 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 $\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1}) \geq 1$. $\mathbb{I}_{s_1}^{s_2} = \{i \in \mathbb{Z} \mid s_1 \leq i \leq s_2\}$ denotes an index set. The vector containing the signs of a vector $x \in \mathbb{R}^n$ (in terms of values ± 1) is denoted by $\mathrm{Sgn}(x)$, while the support of x is denoted by $\mathrm{Supp}(x) := \{i \in \mathbb{I}_1^n \mid [x]_i \neq 0\}$. For a given $\mathcal{T} \subset \mathbb{I}_1^n$, $x_\mathcal{T}$ denotes the projection of x to the support $\mathcal{T}$, *i.e.*, $[x_\mathcal{T}]_i := [x]_i$ if $i \in \mathcal{T}$ and 0 otherwise.

$X_N \xrightarrow{p} X$ denotes convergence in probability [17]. 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 $N \geq N_0(\varepsilon)$, $P\{m(\varepsilon) < |A_N/B_N| < M(\varepsilon)\} \geq 1 - \varepsilon$. $x_N \in \mathrm{As}\,\mathcal{N}(x_0, P)$ means that the sequence of random variables $\{x_N\}$ converges in distribution to a normal distribution with mean x_0 and covariance P.

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.

III. Problem setup

The most general setup to be considered in this paper is introduced now. Consider the stable discrete-time LTI data-generating system

$$y_t = \frac{B_\mathrm{o}(q)}{A_\mathrm{o}(q)} u_t + \frac{C_\mathrm{o}(q)}{D_\mathrm{o}(q)} e_t, \tag{2}$$

where e_t is a Gaussian white noise sequence of zero mean and variance $\sigma^2 > 0$, u_t is a quasi-stationary signal [2], and

$$A_\mathrm{o}(q) = 1 + a_1^\mathrm{o} q^{-1} + \cdots + a_{n_\mathrm{a}}^\mathrm{o} q^{-n_\mathrm{a}}, \tag{3a}$$

$$B_\mathrm{o}(q) = b_1^\mathrm{o} q^{-1} + \cdots + b_{n_\mathrm{b}}^\mathrm{o} q^{-n_\mathrm{b}}, \tag{3b}$$

$$C_\mathrm{o}(q) = 1 + c_1^\mathrm{o} q^{-1} + \cdots + c_{n_\mathrm{c}}^\mathrm{o} q^{-n_\mathrm{c}}, \tag{3c}$$

$$D_\mathrm{o}(q) = 1 + d_1^\mathrm{o} q^{-1} + \cdots + d_{n_\mathrm{d}}^\mathrm{o} q^{-n_\mathrm{d}}, \tag{3d}$$

with q the time-shift operator, $\theta_\mathrm{o} = [\, a_1^\mathrm{o} \dots a_{n_\mathrm{a}}^\mathrm{o} \ b_1^\mathrm{o} \dots b_{n_\mathrm{b}}^\mathrm{o} \,]$ and $\eta_\mathrm{o} = [\, c_1^\mathrm{o} \dots c_{n_\mathrm{c}}^\mathrm{o} \ \ d_1^\mathrm{o} \dots d_{n_\mathrm{d}}^\mathrm{o} \,]$. Due to physical insights or simply to the generality of the representation, we assume as prior knowledge that only few of the parameters θ_o are actually non-zero. Note that for notational convenience, no feedthrough term is assumed. Based on measurements $\mathcal{D}_N := \{u_t, y_t\}_{t=1}^N$, our goal is to estimate a model of this system in the form

$$y_t = \frac{B(q)}{A(q)} u_t + \frac{C(q)}{D(q)} \epsilon_t, \tag{4}$$

where

$$A(q) = 1 + a_1 q^{-1} + \cdots + a_{n_{\mathrm{a}}} q^{-n_{\mathrm{a}}}, \tag{5a}$$
$$B(q) = b_1 q^{-1} + \cdots + b_{n_{\mathrm{b}}} q^{-n_{\mathrm{b}}}, \tag{5b}$$
$$C(q) = 1 + c_1 q^{-1} + \cdots + c_{n_{\mathrm{c}}} q^{-n_{\mathrm{c}}}, \tag{5c}$$
$$D(q) = 1 + d_1 q^{-1} + \cdots + d_{n_{\mathrm{d}}} q^{-n_{\mathrm{d}}}. \tag{5d}$$

In this paper, we assume that the model structure (4) contains the true system (2), *i.e.*, there is no undermodelling.

As an intermediate step in the development of a general sparse estimation procedure for rational model structures (4), we will first concentrate on systems and model structures where $D(q) = A(q)$, $D_{\mathrm{o}}(q) = A_{\mathrm{o}}(q)$ and $C(q) = C_{\mathrm{o}}(q) = 1$. For these restricted model structures, the *Maximum Likelihood* (ML) and the PEM methods are equivalent to a linear regression problem [2].

IV. Methods

In this section, we propose a method for the estimation of model structure (4) which takes into account (possible) sparsity in the parameter vector. To this end, first we present a new method called SPARSEVA (a sparse LASSO-type estimator) for linear regression problems, originally proposed in [14], which provides automatic tuning of its regularization parameters. Next, we will revisit the Steiglitz-McBride method, a well established iterative technique for the estimation of OE model structures, based on the solution of a series of linear regression problems. Finally, we show how to combine these two procedures, the SPARSEVA and Steiglitz-McBride methods, in order to estimate general sparse rational model structures.

A. SPARSEVA

1) Assumptions: As mentioned at the end of Section III, we will first focus on model structures that can be written as a linear regression. Assume that the data is generated by

$$Y_N = \varPhi_N \theta^{\mathrm{o}} + E_N, \tag{6}$$

where $\theta^{\mathrm{o}} \in \mathbb{R}^{n_{\mathrm{g}}}$, $E_N \sim \mathcal{N}(0, \sigma^2 I_N)$ (with $\sigma^2 > 0$), $\varPhi_N \in \mathbb{R}^{N \times n_{\mathrm{g}}}$ and $Y_N \in \mathbb{R}^N$. Furthermore, $\mathcal{T}_{\mathrm{o}} := \mathrm{Supp}(\theta^{\mathrm{o}})$ where $\mathcal{T}_{\mathrm{o}}$ contains the indexes (positions) of the non-zero elements of θ^{o}, while $\bar{\mathcal{T}}_{\mathrm{o}} := \mathbb{I}_1^{n_{\mathrm{g}}} \setminus \mathcal{T}_{\mathrm{o}}$ denotes the positions of the zeros. The model chosen to capture the dynamics of (6) is

$$Y_N = \varPhi_N \theta + E_N, \tag{7}$$

where $\theta \in \mathbb{R}^{n_{\mathrm{g}}}$ is unknown (which also means that $\mathcal{T}_{\mathrm{o}}$ is a priori unknown). With respect to $\varPhi_N$ and its relation to E_N, we will assume that:

1) $N^{-1} \varPhi_N^\top \varPhi_N \xrightarrow{p} \varGamma \succ 0$ as $N \to \infty$,
2) $V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N) \xrightarrow{p} \sigma^2$ as $N \to \infty$, and
3) $\sqrt{N}(\hat{\theta}_N^{\mathrm{LS}} - \theta^{\mathrm{o}}) \in \mathrm{As}\, \mathcal{N}(0, \sigma^2 M)$, where M is a non-singular matrix.

Here,

$$V(\theta, \mathcal{D}_N) := \frac{1}{N}(Y_N - \varPhi_N \theta)^\top (Y_N - \varPhi_N \theta), \tag{8}$$

is the least-squares cost (ℓ_2-loss of the prediction), and

$$\hat{\theta}_N^{\mathrm{LS}} := \left(\varPhi_N^\top \varPhi_N\right)^{-1} \varPhi_N^\top Y_N, \tag{9}$$

is the least-squares estimate.

A particular case of interest for us is the ARX model structure

$$A(q) y_t = B(q) u_t + \epsilon_t, \tag{10}$$

which corresponds to structure (4) with $C(q) = 1$ and $D(q) = A(q)$. This ARX structure can be written in the linear regression form (7) where $Y_N := [\, y_{n_{\mathrm{a}}+1} \ldots y_N \,]^\top$, $E_N := [\, \epsilon_{n_{\mathrm{a}}+1} \ldots \epsilon_N \,]^\top$, $\theta := [\, a_1 \ldots a_{n_{\mathrm{a}}} \; b_1 \ldots b_{n_{\mathrm{b}}} \,]^\top$ and

$$\varPhi_N = \begin{bmatrix} -y_{n_{\mathrm{a}}} & \cdots & -y_1 & u_{n_{\mathrm{a}}} & \cdots & u_{n_{\mathrm{a}}-n_{\mathrm{b}}+1} \\ \vdots & & \vdots & \vdots & & \vdots \\ -y_{N-1} & \cdots & -y_{N-n_{\mathrm{a}}} & u_{N-1} & \cdots & u_{N-n_{\mathrm{b}}} \end{bmatrix}. \tag{11}$$

(For the sake of simplicity, we assume that $n_{\mathrm{a}} \geq n_{\mathrm{b}}$.)

Remark 4.1: Assumptions 1-3 are not necessary conditions in order to obtain estimates with nice statistical properties (as seen in Section V). For example, in case $\sigma = 0$, i.e., the data is noiseless, then the least squares estimate will be exactly equal to θ^{o} for $N \geq n_{\mathrm{g}}$. The case where $\sigma^2 > 0$ is certainly more interesting and practically relevant for system identification. Depending on the particular model structure considered, general sufficient conditions for Assumptions 1-3 to hold are global identifiability of the model structure and persistence of excitation of the input signal [2].

Remark 4.2: Notice that Assumptions 1-3 also hold if $\varPhi_N$ is deterministic and satisfies $N^{-1} \varPhi_N^\top \varPhi_N \to \varGamma > 0$ as $N \to \infty$.

2) Method: The method we propose for estimating a sparse θ is based on the following steps:

i) Compute the ordinary least-squares estimate $\hat{\theta}_N^{\mathrm{LS}}$ via (9).

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

$$\min_{\theta \in \mathbb{R}^{n_{\mathrm{g}}}} \quad \|\theta\|_1, \tag{12a}$$
$$\text{s.t.} \quad V(\theta, \mathcal{D}_N) \leq V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N)(1 + \varepsilon_N), \tag{12b}$$

where $\varepsilon_N > 0$ and $n_{\mathrm{g}} := n_{\mathrm{a}} + n_{\mathrm{b}}$. 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, let $\mathcal{T}$ correspond to the indexes of the non-zero elements in $\hat{\theta}_N$. Define $\varPhi_{N,\mathcal{T}}$ to be the matrix formed from the columns of $\varPhi_N$ listed in $\mathcal{T}$ and then compute the least-squares estimate of a θ of reduced dimension based on the model (7) with $\varPhi_{N,\mathcal{T}}$. 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*), and the estimate is denoted as $\hat{\theta}_N$. When Step iii) is also used, we call the method SPARSEVA-RE, indicating that the non-zero parameters are re-estimated (using least-squares); the corresponding estimate is denoted $\hat{\theta}_N^{\mathrm{RE}}$.

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

$$\min_{\theta \in \mathbb{R}^{n_{\mathrm{g}}}} \quad \|w_N \odot \theta\|_1, \tag{13a}$$
$$\text{s.t.} \quad V(\theta, \mathcal{D}_N) \leq V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N)(1 + \varepsilon_N), \tag{13b}$$

where $w_N \in \mathbb{R}_+^{n_{\mathrm{g}}}$ is given by $[w_N]_i := 1/|[\hat{\theta}_N^{\mathrm{LS}}]_i|^\gamma$ with $i \in \mathbb{I}_1^{n_{\mathrm{g}}}$ and $\gamma > 0$ being an arbitrary constant. We denote the method obtained from Step i) and (13) by A-SPARSEVA (*Adaptive SPARSEVA*) and the corresponding estimate by $\hat{\theta}_N^{\mathrm{A}}$; the method with all three steps, in this case, is denoted as A-SPARSEVA-RE and the corresponding estimate is $\hat{\theta}_N^{\mathrm{A-RE}}$.

This adaptive version is inspired by the adaptive LASSO [18]. The proposed estimation scheme is summarized in terms of Algorithm 1. The specific choice of ε_N is discussed later. It is important to highlight that both (12) and (13) are convex for linear regression problems.

Algorithm 1 A-SPARSEVA-RE

```
Require: a data record $\mathcal{D}_N = \{u_t, y_t\}_{t=1}^N$ of (2) and the
    model structure (10) characterized by the parameters
    $\theta = [a_1 \dots b_{n_\mathrm{b}}]^\top \in \Theta \subseteq \mathbb{R}^{n_\mathrm{g}}$. Assume that $\mathcal{D}_N$ is
    informative w.r.t. (10), see [2].
1: Compute $\hat{\theta}_N^{\mathrm{LS}}$ via (9).
2: Set $\varepsilon_N = 2n_\mathrm{g}/N$ (or $\varepsilon_N = (n_\mathrm{g} \log N)/N$) and compute
   $V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N)$.
3: Obtain the sparse estimate $\hat{\theta}_N^{\mathrm{A}}$ by solving (13). (In the
   non-adaptive case, solve (12) to obtain $\hat{\theta}_N$)
4: Based on a threshold $0 \leq \varepsilon_* \ll 1$, select a minimal $\mathcal{T} \subseteq$
   $\mathrm{Supp}(\hat{\theta}_N^{\mathrm{A}})$ such that $\|\hat{\theta}_{N,\mathcal{T}}^{\mathrm{A}} - \hat{\theta}_N^{\mathrm{A}}\|_1 \leq \varepsilon_* \|\hat{\theta}_N^{\mathrm{A}}\|_1$.
5: Estimate $\theta_N^{\mathrm{A-RE}}$ via (9) with $\Phi_{N,\mathcal{T}}$.
6: return estimated model (10).
```

3) Discussion of the method: The idea behind SPARSEVA is based on *Akaike's Information Criterion* (AIC). Let $\mathcal{D}_N^{\mathrm{val}}$ denote a fresh validation data set (corresponding to a different realization of Φ_N and the noise E_N). Then, for linear regression problems, c.f. [2], it is easily shown that

$$\mathrm{E}_{\mathrm{val}}\left\{\mathrm{E}_{\mathrm{est}}\left\{V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N^{\mathrm{val}})\right\}\right\} = \left(1 + \frac{2n_\mathrm{g}}{N}\right) \mathrm{E}_{\mathrm{est}}\left\{V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N)\right\} \tag{14}$$

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

The relation (14) suggests that a way to perform model selection without using a validation data set is to minimize

$$\left(1 + \frac{2n_\mathrm{g}}{N}\right) V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N), \tag{15}$$

with respect to n_g, the number of estimated parameters. This is Akaike's AIC (or Final Prediction Error, FPE) criterion for model selection.

In view of this, with the choice $\varepsilon_N = 2n_\mathrm{g}/N$, (12) can be seen as a way to estimate a sparse (due to the ℓ_1-norm) model such that its performance is similar to $V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N^{\mathrm{val}})$. Thus, unlike for the LASSO, there is a natural choice of the "regularization" parameter ε_N for SPARSEVA, which corresponds to a particular level set of the ℓ_2-loss function (8), in which the loss of the sparse solution is expected to lay. In the LASSO case, the ℓ_2 cost of the prediction error is minimized for a given sparsity level, *i.e.*, $\|\theta\|_1 < \varepsilon$, see (1). As the ℓ_1 norm of the optimal estimate for θ is generally unknown, it is much harder in practice to develop a selection scheme for ε in the LASSO method. This is the motivation for introducing (12).

It should be noted that the convex optimization program

$$\min_{\theta} \quad \|\theta\|_1, \tag{16a}$$

$$\text{s.t.} \quad V(\theta, \mathcal{D}_N) \leq \varepsilon, \tag{16b}$$

has been used before for signal recovery in the compressive sensing context [19, 20], *i.e.*, when the number of observations N is less than the number of estimated parameters n_g. Our contribution lies in the suggestion to use ε_N according to (12), in particular with ε_N chosen by the AIC rule $\varepsilon_N = 2n_\mathrm{g}/N$, and in the adaptive version (13) inspired by [18].

The use of a threshold ε_* in Step 4 of Algorithm 1 to determine the support of $\hat{\theta}_N^A$ is considered merely for numerical purposes: many numerical methods for solving (12) or (13) (e.g., CVX [21]) deliver a solution $\hat{\theta}_N^A$ which is sparse only up to some numerical precision (e.g., 10^{-10}). The choice of ε_* should be made according to the precision of method used to solve (12) or (13) (typically an order of magnitude larger than such tolerance). In practice, since such tolerances are much smaller than the achievable statistical accuracy, the effect of thresholding with ε_* is negligible[1] (in statistical terms). Notice that for the theoretical results of Section V we assume infinite numerical precision, hence we take $\varepsilon_* = 0$.

B. Steiglitz-McBride Method

Consider now an *Output-Error* (OE) model structure,

$$y_t = \frac{B(q)}{A(q)} u_t + \epsilon_t, \tag{17}$$

which corresponds to (4) with $C(q) = D(q) = 1$. It is well known, see [2], that the least-squares estimator $\hat{\theta}_N^{\mathrm{LS}} := (\Phi_N^\top \Phi_N)^{-1} \Phi_N^\top Y_N$ (where Φ_N is given as in (11)) is biased, and the cost function of PEM for this model structure is non convex, hence its minimization is difficult and may suffer from local minima.

One technique for estimating models of type (17) from least-squares estimates is the so-called Steiglitz-McBride method [22]. The idea of this method is to iteratively pre-filter u_t and y_t by $1/\hat{A}^{(k)}(q)$ resulting in the filtered data set $\mathcal{D}_N^{(k)}$, where $\hat{A}^{(k)}(q)$ is an estimate of the $A(q)$ polynomial (at step k). Then, least-squares estimation is applied on $\mathcal{D}_N^{(k)}$, assuming a model structure such as (10), which gives estimates $\hat{A}^{(k+1)}(q)$ and $\hat{B}^{(k+1)}(q)$. This procedure is usually initialized by taking $\hat{A}^{(0)}(q) = 1$, and stopped when the estimates $\hat{A}^{(k)}(q)$ and $\hat{B}^{(k)}(q)$ do not change much from one iteration to the next.

The Steiglitz-McBride algorithm has been extensively studied in the literature [23, 24]. In particular, it is known to give unbiased estimates only if the true system belongs to an OE structure (17), and its global convergence properties are still largely an open problem. In addition, the Steiglitz-McBride is not asymptotically efficient for (17).

In [16], an interesting extension of the Steiglitz-McBride algorithm has been developed, which gives consistent estimates even for Box-Jenkins model structures (4). This extension is based on a preliminary step, where a high order ARX model

$$A_{\mathrm{HO}}(q) y_t = B_{\mathrm{HO}}(q) u_t + \epsilon_t, \tag{18}$$

with

$$A_{\mathrm{HO}}(q) = 1 + a_1^{\mathrm{HO}} q^{-1} + \cdots + a_m^{\mathrm{HO}} q^{-m}, \tag{19a}$$

$$B_{\mathrm{HO}}(q) = b_1^{\mathrm{HO}} q^{-1} + \cdots + b_m^{\mathrm{HO}} q^{-m}, \tag{19b}$$

is fitted to $\mathcal{D}_N$, and used then to pre-filter the data, *i.e.*, to generate the signals

$$y_t^{\mathrm{F}} := \hat{A}_{\mathrm{HO}}(q) y_t, \quad u_t^{\mathrm{F}} := \hat{A}_{\mathrm{HO}}(q) u_t.$$

[1] Many sparse estimation methods rely in practice on a final thresholding step for support set recovery (as Step 4 in Algorithm 1), sometimes even without explicitly saying so.

This filtered data is then used in place of the original input and output signals of (17) on which the Steiglitz-McBride procedure is executed, resulting in estimates of the polynomials $A(q)$ and $B(q)$. The intuition behind this method is that $1/\hat{A}_{\mathrm{HO}}(q)$ should be a reasonable estimate of the noise model $C_{\mathrm{o}}(q)/D_{\mathrm{o}}(q)$, hence the pre-filtering stage should "whiten" the noise (as seen from the output). This means that the standard Steiglitz-McBride method could then deliver a consistent estimate of the polynomials $A(q)$ and $B(q)$.

An important issue regarding the Steiglitz-McBride method is the stability of the pre-filters $\hat{A}^{(k)}$ and A_{HO}: these filters are not guaranteed to be stable, so if at some iteration the estimated filter is unstable, the Steiglitz-McBride method cannot continue. One way to overcome this issue is to split the unstable filter $\hat{A}^{(k)}$ as $\hat{A}_{+}^{(k)}\hat{A}_{-}^{(k)}$, where $\hat{A}_{+}^{(k)}$ is stable and $\hat{A}_{-}^{(k)}$ is anti-stable (the constant factor can be arbitrarily assigned to any of these factors), pre-filter the data forward in time using $1/\hat{A}_{+}^{(k)}$, and then use the filter $1/\hat{A}_{-}^{(k)}$ *backwards in time*. This technique has been used in other contexts within system identification (e.g., [25]) and it preserves the second-order properties of the Steiglitz-McBride method (since it corresponds to the time-domain equivalent of the Sanathanan-Koerner method [26]).

Some results on the accuracy of the extended Steiglitz-McBride method are detailed in Section V.

C. Estimation of Sparse Output-Error Models

As mentioned in Section I, SPARSEVA and ℓ_1-penalized estimators cannot be directly applied to model structures such as (4), because the PEM cost function is non convex. However, techniques such as Steiglitz-McBride, which rely on least-squares optimization, can be directly extended to use ℓ_1-penalized estimators in order to deliver sparse models.

Based on the previous discussion, Algorithm 2 provides estimation of sparse rational OE models (17).

Remark 4.3: Note that in Step 8, A-SPARSEVA can be used to impose several different sparsity patterns on the $A(q)$ and $B(q)$ polynomials. For example, if we only want to impose sparsity on $A(q)$, then the ℓ_1-norm in the cost function of (13) can be modified so that only the coefficients of $A(q)$ are included.

Remark 4.4: Based on validation data, optimization of ε_N can also be applied to recover the exact sparsity structure of θ. However, re-optimizing such quantity (using *e.g.* cross-validation) is equivalent to optimizing for the regularization parameter in a standard LASSO estimator (inclusion of $V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N)$ in (12b) is not necessary). This might refine the results for relatively small data-lengths N under considerable noise, but at the expense of a much higher computational load. Hence a clearly important feature of the proposed SPARSEVA scheme is an automatic choice of ε_N guaranteeing a reliable performance.

V. Main results

In this section, we present the main technical results about the asymptotic properties of the introduced methods. For theoretical purposes, we neglect numerical errors due to finite precision, hence we assume for simplicity that $\varepsilon^* = 0$.

Algorithm 2 OE-SPARSEVA with Steiglitz-McBride

```
Require: a data record $\mathcal{D}_N = \{u_t, y_t\}_{t=1}^N$ of (2) and the model structure (17) characterized by the parameters $\theta = [a_1 \dots b_{n_\mathrm{b}}]^\top \in \Theta \subseteq \mathbb{R}^{n_\mathrm{a}+n_\mathrm{b}}$. Assume that $\mathcal{D}_N$ is informative w.r.t. (17) and (17) is globally identifiable on $\Theta$ [2].
1: Let $m \gg n_\mathrm{a}$ and fit using least-squares the high order ARX model described by (18) to the measurements $\mathcal{D}_N$, resulting in $\hat{A}_{\mathrm{HO}}(q)$ and $\hat{B}_{\mathrm{HO}}(q)$.
2: Filter the data $\mathcal{D}_N$ as
       $y_t^{\mathrm{F}} := \hat{A}_{\mathrm{HO}}(q) y_t, \quad u_t^{\mathrm{F}} := \hat{A}_{\mathrm{HO}}(q) u_t.$
3: Set $k = 0$, and let $\hat{A}^{(0)}(q) = 1$, $\hat{B}^{(0)}(q) = 0$ and consequently $\hat{\theta}_N^{(0)} = 0$.
4: repeat
5:     $k \leftarrow k+1$ and filter the data $\mathcal{D}_N^{\mathrm{F}} = \{u_t^{\mathrm{F}}, y_t^{\mathrm{F}}\}_{t=1}^N$ as
           $y_t^{\mathrm{F}(k)} := \frac{1}{\hat{A}^{(k-1)}(q)} y_t^{\mathrm{F}}, \quad u_t^{\mathrm{F}(k)} := \frac{1}{\hat{A}^{(k-1)}(q)} u_t^{\mathrm{F}}.$
6:     Fit, using least-squares, a model of the form
           $A^{(k)}(q) y_t^{\mathrm{F}(k)} = B^{(k)}(q) u_t^{\mathrm{F}(k)} + \epsilon_t^{(k)}, \qquad (20)$
       resulting in the estimates $\hat{A}^{(k)}$, $\hat{B}^{(k)}$ and the associated parameter vector $\hat{\theta}_N^{(k)}$.
7: until $\hat{\theta}_N^{(k)}$ has converged or the maximum number of iterations is reached.
8: Apply A-SPARSEVA (with least-squares re-estimation) to the model
       $A(q) y_t^{\mathrm{F}(k+1)} = B(q) u_t^{\mathrm{F}(k+1)} + \epsilon_t^{(k+1)}. \qquad (21)$
9: return estimated model (17).
```

A. SPARSEVA

We will first investigate the theoretical properties of SPARSEVA and its adaptive variant.

1) Consistency: Regarding consistency of the estimator w.r.t. (6), we have the following results:

Theorem 5.1 (Consistency of (A-)SPARSEVA): Under the assumptions of Section IV-A1, and $\theta^{\mathrm{o}} \neq 0$, the SPARSEVA and A-SPARSEVA estimators are consistent in probability (*i.e.*[2], $\hat{\theta}_N^{(\mathrm{A})} \xrightarrow{p} \theta^{\mathrm{o}}$) if and only if $\varepsilon_N \to 0$. In particular, $\|\hat{\theta}_N^{(\mathrm{A})} - \theta^{\mathrm{o}}\|_2 = O_p(N^{-1/2} + \sqrt{\varepsilon_N})$ uniformly in θ^{o}.

Proof: See Appendix B. ■

Corollary 5.1 (Exact order of consistency): Subject to the assumptions of Theorem 5.1, if $\varepsilon_N \to 0$, but $N\varepsilon_N \to \infty$, then $\|\hat{\theta}_N^{(\mathrm{A})} - \theta^{\mathrm{o}}\|_2 \asymp_p \sqrt{\varepsilon_N}$ (*c.f.* Section II for the definition of $\asymp_p$).

Proof: See Appendix C. ■

2) Sparseness: Since $V(\theta, \mathcal{D}_N)$ is quadratic, the constraint (12b) corresponds to an ellipsoid in $\Theta \subseteq \mathbb{R}^{n_\mathrm{g}}$. The solution to (12a) 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

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

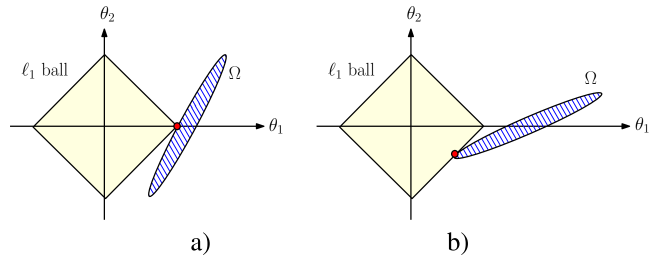


Fig. 1. The geometry of (12). In a), a sparse solution ($\theta_2 = 0$) is obtained but not in b).

Figure 1.b, the solution will not be sparse. The shape of the ellipsoid is determined by the regressor matrix $\varPhi_N$.

Various measures to ensure sparsity have been suggested, *e.g.* [27, 28]. The adaptive SPARSEVA (13) is inspired by [18]. We now establish the exact conditions on ε_N for the adaptive SPARSEVA to generate sparse estimates (recovery of the true support of $\theta^{\rm o}$).

Theorem 5.2 (Sparseness of the adaptive SPARSEVA): Under the assumptions of Section IV-A1 together with $\varepsilon_N \to 0$ and $\theta^{\rm o} \neq 0$, A-SPARSEVA (13) satisfies the sparseness property (*i.e.*, $P\{\mathrm{Supp}(\hat{\theta}_N^{\rm A}) = \mathrm{Supp}(\theta^{\rm o})\} \to 1$ if $N\varepsilon_N \to \infty$). If $N\varepsilon_N \to \infty$ does not hold, then A-SPARSEVA does not have the sparseness property.

Proof: See Appendix C. ■

Remark 5.1: It can be shown that when the regressors are orthonormal, *i.e.*, $N^{-1}\varPhi_N^\top \varPhi_N = I$, then Theorem 5.2 holds also for SPARSEVA. ■

3) Adaptive SPARSEVA and the Oracle property: From the preceding 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 5.1, such a choice of ε_N gives rise to a non efficient estimator (since the order of convergence of $\hat{\theta}_N^{\rm A}$ to $\theta^{\rm o}$ 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 IV) 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 5.3 (The Oracle property): Consider the assumptions in Theorem 5.2 and that $N\varepsilon_N \to \infty$. Then,

$$\sqrt{N}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \in \mathrm{As}\,\mathcal{N}(0, M^{-1}),$$

where M is the information matrix when it is known which elements of $\theta^{\rm o}$ are zero.

Proof: See Appendix D. ■

Remark 5.2: We remark that it is clear from the proof of Theorem 5.3 that such result holds if we replace the use of $\hat{\theta}_N^{\rm A}$ as an estimator of the location of the non-zero components of $\theta^{\rm o}$ by any other $\sqrt{N}$-consistent estimator of such components. For example, Remark 5.1 implies that Theorem 5.3 holds for SPARSEVA-RE when the regressors are orthonormal. ■

4) Minimax rate optimality: Say that we are interested in an estimate $\hat{\theta}_N$ of $\theta^{\rm o}$ such that the *risk* $R(\hat{\theta}_N, \theta^{\rm o}) := \mathrm{E}\{\|\hat{\theta}_N - \theta^{\rm o}\|_2^2\}$ is small. Since the risk R depends on the unknown true value $\theta^{\rm o}$, it is relevant to study the worst-case performance, $\sup_{\theta^{\rm o} \in \mathbb{R}^{n_g}} R(\hat{\theta}_N, \theta^{\rm o})$. In particular, we will focus on the rate of decay of R with respect to the number of samples N. The following definition is appropriate:

Definition 5.1 (Minimax rate optimality): The estimator $\hat{\theta}_N$ is *minimax rate optimal* over the class of all estimators of $\theta^{\rm o}$, if $\sup_{\theta^{\rm o} \in \mathbb{R}^{n_{\rm g}}} R(\hat{\theta}_N, \theta^{\rm o})$ converges to zero at the same rate as $\inf_{\hat{\delta} \in \mathbb{R}^{n_{\rm g}}} \sup_{\theta^{\rm o} \in \mathbb{R}^{n_{\rm g}}} R(\hat{\delta}, \theta^{\rm o})$, where $\hat{\delta}$ ranges over all estimators of $\theta^{\rm o}$ based on the observation vector Y_N.

The conditions for adaptive SPARSEVA and its re-estimated version to be minimax rate optimal are as follows:

Theorem 5.4 (Minimax rate optimality): Under the Assumptions IV-A1, A-SPARSEVA (13) and $\hat{\theta}_N^{\rm A-RE}$ are minimax rate optimal if and only if $\varepsilon_N = O_p(N^{-1})$.

Proof: See Appendix E. ■

Remark 5.3: Theorem 5.4 shows that A-SPARSEVA and A-SPARSEVA-RE cannot be minimax-rate-optimal and have the oracle property at the same time. This fundamental tradeoff is present in all model selection procedures, as shown in [29, 30].

Remark 5.4: The oracle property (Theorem 5.3) seems to contradict the Cramér-Rao inequality, according to which the covariance of an unbiased estimator (which does not assume the sparsity structure of the parameter being estimated) cannot be smaller than the inverse of the full Fisher information matrix (which does not assume such sparsity pattern). In fact, there is no apparent contradiction: all sparse estimators are indeed "super-efficient", in the sense that they can beat the Cramér-Rao bound (when they are tuned to enjoy the oracle property). The reason is that these estimators are not unbiased, but only asymptotically unbiased, and they rely on non-smooth functions, such as the ℓ_1 norm, so the conditions for the Cramér-Rao inequality do not hold for these estimators. This is a well known issue (see [29]), as sparse estimators can be seen as a combination of model structure selection and estimation (or "pre-test estimators"), resembling Hodges-type super-efficient estimators. This, of course, does not come for free: as seen in the previous remark, if a sparse estimator is tuned to satisfy the oracle property, it looses its minimax rate optimality.

Remark 5.5: Notice that the scaling of the parameters in $\theta_{\rm o}$ does not seem to play a major role in the estimation performance of Algorithm 2, at least asymptotically in N, since A-SPARSEVA weights the ℓ_1 norm by the inverse of the estimates in $\hat{\theta}_N^{\rm LS}$, which compensates for the relative size of the components of $\theta^{\rm o}$.

Remark 5.6: As seen in the theorems of this section, the consistency, sparseness, oracle and minimax-rate-optimality properties do not depend on constant factors in ε_N, but only on its asymptotic rate as a function of N. This comes from the fact that the described properties are asymptotic in nature, which means that constant factors may affect the finite sample behavior of the estimator, but their effect becomes negligible for large N. The irrelevance of constant factors is also common in the consistency of standard model selection criteria; see, e.g., [3, Section 11.5].

B. Steiglitz-McBride method

The modified Steiglitz-McBride method presented in this paper, which includes a stabilization scheme (based on reflecting the unstable poles of the prefilter) and a high order ARX pre filtering step, is due to Y. Zhu [16]. This method, as well as the original Steiglitz-McBride algorithm, can be expected to be globally convergent if the signal to noise ratio is sufficiently high (*c.f.* [23]), but its global convergence properties in the general case are not well understood yet. However, preliminary results seem to indicate that the equilibrium point of the modified method is a consistent and asymptotically efficient estimator of $A_{\mathrm{o}}(q)$ and $B_{\mathrm{o}}(q)$ for general Box-Jenkins model structures[3] (4).

C. OE-SPARSEVA

The combination of A-SPARSEVA and the modified Steiglitz-McBride method, OE-SPARSEVA, as presented in Section IV-C, can be expected to have attractive asymptotic properties. Indeed, by combining the theoretical results of its components, we obtain the following result:

Theorem 5.5 (Properties of OE-SPARSEVA): Under the assumptions of Section III, OE-SPARSEVA (assuming convergence of the Steiglitz-McBride iterations)

1) is consistent in probability if and only if $\varepsilon_N \to 0$,
2) has the sparseness property, for $\varepsilon_N \to 0$ and $\theta^{\mathrm{o}} \neq 0$, if and only if $N\varepsilon_N \to \infty$,
3) has the oracle property if $\varepsilon_N \to 0$ and $N\varepsilon_N \to \infty$.

Proof: See Appendix F. ■

D. Equivalence with other sparse estimators

Next, we show how the introduced A-SPARSEVA estimator is related to the LASSO and the NNG (resembling the proof of the duality result in [31, Theorem 3]). First, consider the adaptive version of the LASSO estimator (1) (for $\gamma = 1$), which was first introduced in [18]:

$$\min_{\theta\in\mathbb{R}^{n_{\mathrm{g}}}} \quad V(\theta,\mathcal{D}_N) \tag{22a}$$

$$\text{s.t.} \quad \|w_N \odot \theta\|_1 \leq \varepsilon_{\mathrm{L}}. \tag{22b}$$

This estimator can be written in the *Lagrangian* form

$$\mathit{\Lambda}_{\mathrm{L}}(\theta,\lambda_{\mathrm{L}}) = V(\theta,\mathcal{D}_N) + \lambda_{\mathrm{L}}\big(\|w_N \odot \theta\|_1 - \varepsilon_{\mathrm{L}}\big), \tag{23}$$

with $\lambda_{\mathrm{L}} \geq 0$. The optimum of (22) is obtained at the optimum of

$$\max_{\lambda_{\mathrm{L}}\geq 0}\min_{\theta\in\mathbb{R}^{n_{\mathrm{g}}}} \mathit{\Lambda}_{\mathrm{L}}(\theta,\lambda_{\mathrm{L}}). \tag{24}$$

Similarly, (13) has the Lagrangian:

$$\begin{aligned}\mathit{\Lambda}_{\mathrm{S}}(\theta,\lambda_{\mathrm{S}}) = {}& \|w_N \odot \theta\|_1 + \\ & \lambda_{\mathrm{S}}\big(V(\theta,\mathcal{D}_N) - V(\hat{\theta}_N^{\mathrm{LS}},\mathcal{D}_N)(1+\varepsilon_N)\big),\end{aligned} \tag{25}$$

Notice that both $V(\centerdot,\mathcal{D}_N)$ and $\|\centerdot\|_1$ are convex functions, $V(\centerdot,\mathcal{D}_N)$ is strongly convex and all constraints satisfy the constraint qualification, hence solutions of (24) and (25), i.e., $(\hat{\theta}_{\mathrm{L}}(\epsilon_{\mathrm{L}}),\lambda_{\mathrm{L}}^*(\epsilon_{\mathrm{L}}))$ and $(\hat{\theta}_{\mathrm{S}}(\varepsilon_{\mathrm{S}}),\lambda_{\mathrm{S}}^*(\varepsilon_{\mathrm{S}}))$ are unique with no duality gap.

For a given ε_{L}, let ε_N be such that $V(\hat{\theta}_N^{\mathrm{LS}},\mathcal{D}_N)(1+\varepsilon_N)$ is equal to the minimum value of (22). For this choice of ε_N, the feasibility sets $U_{\mathrm{L}} := \{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : \|w_N \odot \theta\|_1 \leq \varepsilon_{\mathrm{L}}\}$ and $U_{\mathrm{S}} := \{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : V(\theta,\mathcal{D}_N) \leq V(\hat{\theta}_N^{\mathrm{LS}},\mathcal{D}_N)(1+\varepsilon_N)\}$ are convex and intersect at exactly one point[4], $\hat{\theta}_{\mathrm{L}}(\varepsilon_{\mathrm{L}}) = \hat{\theta}_{\mathrm{S}}(\varepsilon_{\mathrm{N}})$ (*c.f.* Figure 1). The reverse of this argument also holds respectively. This shows that if $\lambda_{\mathrm{S}}^* \neq 0$ and $\lambda_{\mathrm{L}}^* \neq 0$, then there is a bijective relation between ε_{L} and ε_N such that $\theta^* := \hat{\theta}_{\mathrm{S}}(\varepsilon_N) = \hat{\theta}_{\mathrm{L}}(\varepsilon_{\mathrm{L}})$. Notice, however, that this relation (which is induced by the KKT conditions of (23) and (25)) is dependent on the optimal solution θ^*, i.e., it is data-dependent (center of U_{S} depends on $\hat{\theta}_N^{\mathrm{LS}}$ and its shape depends on $\mathit{\Phi}_N$).

As a next step, consider the NNG in the form of

$$\min_{\bar{w}\in\mathbb{R}^{n_{\mathrm{g}}}} \quad V(\bar{w}\odot\hat{\theta}_N^{\mathrm{LS}},\mathcal{D}_N) + \lambda_{\mathrm{N}}\|\bar{w}\|_1, \tag{26a}$$

$$\text{s.t.} \quad \bar{w} \geq 0, \tag{26b}$$

which provides the parameter estimate $\hat{\theta}_N^{\mathrm{NNG}} := \bar{w}^* \odot \hat{\theta}_N^{\mathrm{LS}}$ with $\bar{w}^*$ being the optimum of (26). Introduce a new variable $\theta = \bar{w}\odot\hat{\theta}_N^{\mathrm{LS}}$ which, if substituted into (26), gives

$$\min_{\theta\in\mathbb{R}^{n_{\mathrm{g}}}} \quad V(\theta,\mathcal{D}_N) + \lambda_{\mathrm{N}}\|w_N\odot\theta\|_1 \tag{27a}$$

$$\text{s.t.} \quad \theta\odot\hat{\theta}_N^{\mathrm{LS}} \geq 0, \tag{27b}$$

as $[w_N]_i := 1/|[\hat{\theta}_N^{\mathrm{LS}}]_i|$. Therefore, by comparing (27) and (24), as observed in [18], we see that the NNG corresponds to the adaptive LASSO with an additional sign constraint (via a suitable, data-dependent, bijection between λ_N and the optimal λ_{L}^*).

Based on the previous derivations, the conclusion is that A-SPARSEVA, the adaptive LASSO and the NNG can be all seen as the same sparse estimator under a specific choice of their regularization (penalization) parameter (and an additional sign constraint for the NNG). This highlights that the real advantage of the A-SPARSEVA scheme is the automatic selection of this parameter, implicitly ensuring either the oracle or the minimax rate optimality properties.

We should emphasize again, however, that the relation between A-SPARSEVA, the adaptive LASSO and the NNG is in general data- (or θ^*-) dependent. This means that, even though their regularization paths are equivalent (modulo monotonic transformations of their regularization parameters), it does not seem possible in general to derive a simple, explicit formula to describe these connections without having first to solve the respective convex optimization problems. In other words, the automatic tuning provided by SPARSEVA cannot be easily translated to the LASSO or NNG formulations.

VI. Numerical example

In this section, we will provide numerical evidence of the performance of the methods developed in Section IV.

A. SPARSEVA

We illustrate the properties and performance of the SPARSEVA approach and compare it with other methods using Example 4.1 in [27]. In this example,

$$\begin{aligned} A_{\mathrm{o}}(q) &= 1, & B_{\mathrm{o}}(q) &= 3q^{-1} + 1.5q^{-2} + 2q^{-5}, \\ C_{\mathrm{o}}(q) &= 1, & D_{\mathrm{o}}(q) &= A_{\mathrm{o}}(q) = 1. \end{aligned}$$

This system has a *Finite Impulse Response* (FIR) structure, which corresponds to a simple regression setup. To identify it from data based on the previously proposed estimation scheme, consider the model structure (10) with $n_{\mathrm{a}} = 0$ and $n_{\mathrm{b}} = 8$, which results in the true parameter vector

$$\theta^{\mathrm{o}} = [3 \;\; 1.5 \;\; 0 \;\; 0 \;\; 2 \;\; 0 \;\; 0 \;\; 0]^\top.$$

[3] Even though it is possible to propose variants of Algorithm 2, where either, *e.g.*, ridge regression or a sparse estimator are used instead of least-squares in Steps 1 or 6, preliminary results show that Zhu's method is already asymptotically efficient when the iterations from Steps 4-7 of OE-SPARSEVA are convergent. This suggests that not much may be gained by considering other variants of Algorithm 2.

[4] If $\theta_{\mathrm{S}} \neq \theta_{\mathrm{L}}$ both belong to $U_{\mathrm{S}} \cap U_{\mathrm{L}}$, then both achieve the minimum of (22). However, $(\theta_{\mathrm{S}}+\theta_{\mathrm{L}})/2$ also belongs to $U_{\mathrm{S}} \cap U_{\mathrm{L}}$ (since it is a convex set), and due to the strong convexity of V, $V([\theta_{\mathrm{S}}+\theta_{\mathrm{L}}]/2,\mathcal{D}_N) < V(\theta_{\mathrm{S}},\mathcal{D}_N)$, contradicting the optimality of θ_{S} and θ_{L}. This means that the optimum of (22) is unique and that $U_{\mathrm{S}} \cap U_{\mathrm{L}}$ is a singleton.

Notice that θ^{o} is rather sparse. For the purpose of identification, estimation and validation data sets have been generated. According to the experimental conditions discussed in [27], each data set has been constructed in terms of a regression matrix $\varPhi_N$ with N independently generated rows where in each row the components are standard normal with a correlation between the i^{th} and the j^{th} terms of $0.5^{|i-j|}$. This corresponds to N number of independent experiments for each row, collected into $\varPhi_N$, which have been conducted on the system with a single output measurement y_t generated by an AR filtered white noise: $u_t - 0.5u_{t-1} = \sqrt{1-0.25}w_t$ with $w_t \sim \mathcal{N}(0,1)$. Under these conditions, 100 estimation and 100 validation data sets have been generated for each "data length" $N \in \{10+10k\}_{k=1}^{10}$ resulting in 11×100 estimation and validation data records. The average *Signal to Noise Ratio*[5] (SNR) has been -3.97dB.

Using these data sets, SPARSEVA is compared to the following methods: LS-ORACLE: least-squares estimate of θ using $\varPhi_{\mathcal{T}_{\mathrm{o}}}$ (prior knowledge of the non-zero parameters). Note that this is the ideal estimator and by the Cramér-Rao lower bound w.r.t. the true support of θ_{o}, no other estimator can perform better. However, it cannot be applied in practice as the optimal model structure is unknown. LASSO-GCV is the LASSO,

$$\min_{\theta \in \mathbb{R}^{n_{\mathrm{g}}}} \; V(\theta, \mathcal{D}_N) + \lambda \|\theta\|_1, \tag{28}$$

where the regularization parameter λ is chosen according to generalized cross-validation [32], *i.e.*, the λ that minimizes

$$V(\hat{\theta}_N, \mathcal{D}_N)/(1-p(\lambda)/N)^2 \tag{29}$$

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

$$p(\lambda) = \mathrm{Tr}\left\{\varPhi_N \left(\varPhi_N^\top \varPhi_N + \lambda W^\dagger\right)^{-1} \varPhi_N^\top\right\}, \tag{30a}$$

$$W = \mathrm{Diag}(|\hat{\theta}_N|), \tag{30b}$$

with $|\cdot|$ taken element-wise. Four variants of SPARSEVA are included: SPARSEVA-AIC/BIC where the constraint ε_N is chosen as AIC ($\varepsilon_N = 2n_{\mathrm{g}}/N$) and BIC ($\varepsilon_N = (n_{\mathrm{g}} \log N)/N$). 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 5.2).

Figure 2 shows the *Mean-Squared Error* (MSE) 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 ε_* 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 MSE of A-SPARSEVA with BIC constraint becomes visually undistinguishable from the MSE of LS-ORACLE[6]; this agrees with the Oracle property, which implies that the difference between these MSE's should vanish asymptotically with N. 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 for almost all other estimators. From Figure 4, 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.

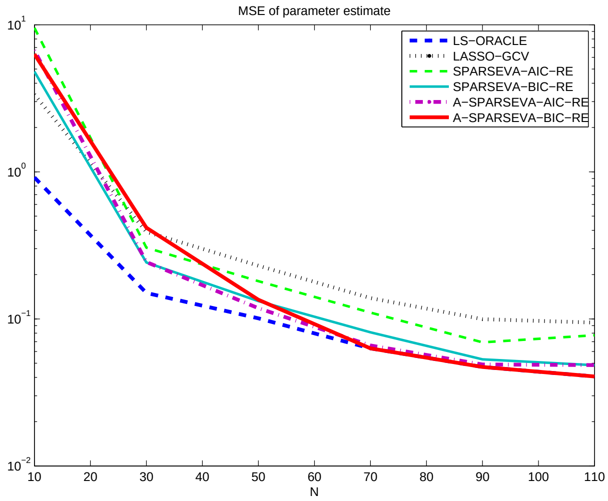


Fig. 2. MSE as a function of the sample size N.

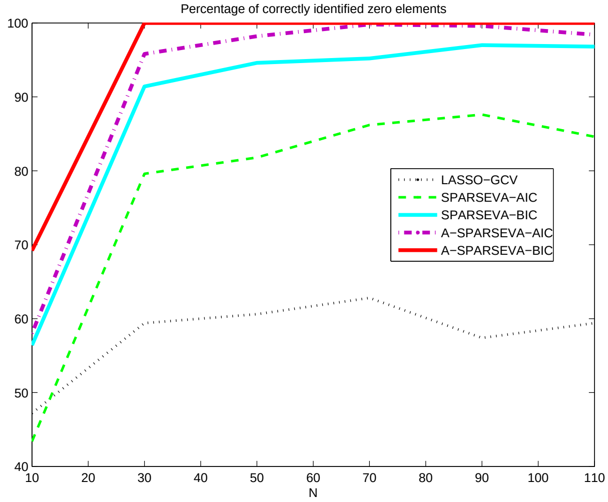


Fig. 3. Percentage of correctly identified zero elements as a function of the sample size N.

B. OE-SPARSEVA

Consider the data-generating system (2) described by the following polynomials:

$$A_{\mathrm{o}}(q) = 1 - 0.1972q^{-2} - 0.2741q^{-8}, \quad B_{\mathrm{o}}(q) = q^{-5} - 8.336q^{-7},$$
$$C_{\mathrm{o}}(q) = 1, \qquad D_{\mathrm{o}}(q) = 1.$$

This system obviously has an OE type of noise structure. To identify this system, consider the model structure (17) with $n_{\mathrm{a}} = 8$ and $n_{\mathrm{b}} = 8$. Even if this corresponds to a

[5] The SNR is defined as $\mathrm{SNR} := 10 \cdot \log_{10}\left(\frac{\|y_t - v_t\|_2^2}{\|v_t\|_2^2}\right)$ where $v_t = \frac{C_{\mathrm{o}}(q)}{D_{\mathrm{o}}(q)} e_t$.

[6] The collapse of the two MSE curves is due to the inherent randomness of the Monte Carlo simulations, not to the equality of the actual MSE of the estimators for $N \geq 70$.

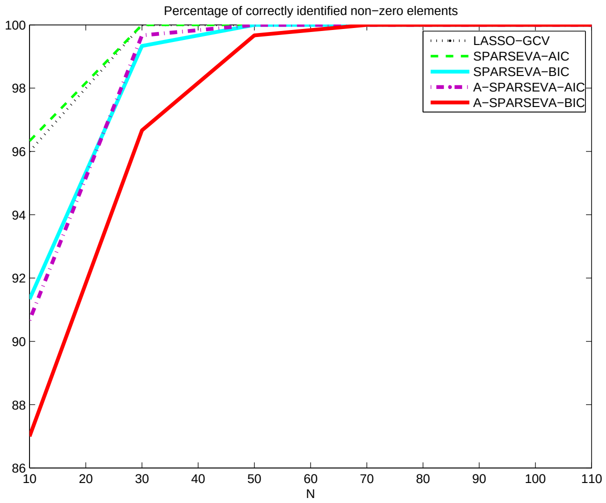


Fig. 4. Percentage of correctly identified non-zero elements as a function of the sample size N.

rather accurate guess of the original order of the polynomials involved, the true parameter vector

$$\theta^{\rm o} = [\,0\ \ -0.1972\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ -0.2741\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0 - 8.3365\,]$$

corresponding to the data-generating system is rather sparse.

Again, a Monte-Carlo study is set up, with 100 estimation and 100 validation data records generated by the system for each data length $N \in \{200 + 50k\}_{k=1}^{37}$, resulting in 37×100 estimation and validation data records with length in the interval $[200, 2000]$. During each computation, u and e have been considered as independent realizations of two white noise sequences with normal distributions $u_t \sim \mathcal{N}(0, 1)$ and $e_t \sim \mathcal{N}(0, \sigma^2)$ respectively. To study the effect of a change in the power of the noise for this case, the generation of the data sequences have been repeated for various standard deviations variances $\sigma \in \{0.0087, 0.275, 1.54, 8.71\}$ corresponding to average SNR's: 30dB, 15dB, 7.5dB, 0dB respectively. This resulted in a total of $4 \times 37 \times 100 = 14800$ estimation and validation data sets defining a serious Monte-Carlo study under various conditions.

Using these data sets, the OE-SPARSEVA described by Algorithm 2, with BIC type of ε_N, LS re-optimization and maximum number of iterations being equal to 50, and the OE algorithm of the *Identification Toolbox* of MATLAB have been applied to estimate the system in the considered model set. In order to fairly assess the quality of the estimates, an SMB-ORACLE estimator in terms of the Steiglitz-McBride method has been also applied with the priori knowledge of which elements of $\theta^{\rm o}$ is zero. The results are compared in terms of

- The MSE of the prediction $\hat{y}_{\hat{\theta}_N}$ on the validation data:

$$\text{MSE} = \frac{1}{N} E\{\|y(k) - \hat{y}_{\hat{\theta}_N}(k)\|_2^2\}, \tag{31}$$

computed as an average over each 100 runs for a given N and σ^2.

- The average of the *fit score* or the *Best Fit Rate* (BFR) [33]:

$$\text{BFR} = 100\% \cdot \max\left(1 - \frac{\|y(k) - \tilde{y}_{\hat{\theta}_N}(k)\|_2}{\|y(k) - \bar{y}\|_2}, 0\right), \tag{32}$$

where $\bar{y}$ is the mean of y and $\tilde{y}_{\hat{\theta}_N}$ is the simulated model output based on the validation data.

- The ℓ_1 parameter estimation error: $\|\hat{\theta} - \theta_{\rm o}\|_1$.
- The percentage of correctly estimated zero elements.

The average results of the 100 Monte-Carlo runs in each cases is given in Figure 5 and the mean and standard deviation of the parameters are given in the SNR= 7.5dB, $N = 2000$ case in Table I. From these results it follows that in the low noise cases (SNR=30dB, 15dB) the proposed OE-SPARSEVA scheme correctly estimates the true support of $\theta_{\rm o}$, *i.e.*, it correctly identifies the underlying model structure of the system and hence it achieves the same results as the SMB-ORACLE approach. The performance difference of the OE approach and the SMB-ORACLE suggests that the reduction of the estimation error can be relatively large by using OE-SPARSEVA in these cases not mentioning the value of really finding which parameters have no role at all in the considered model structure. When the noise increases to a moderate level (SNR=7.5dB), for small data lengths we can observe that OE-SPARSEVA loses the benefits of the regularized optimization scheme by over-estimating the possibly non-zero parameters and achieving worse results than the OE approach. Increasing the number of data points results in a quick recovery of the algorithm and around $N = 800$ it starts achieving similar results as the SMB-ORACLE. We can see that the performance of OE-SPARSEVA asymptotically converges to the SMB-ORACLE approach while the OE has a much slower convergence rate. The same behavior can be observed in the SNR=0dB case. However, initially, the OE approach provides better estimates due to misclassification of the zero elements for this high-noise / low-sample-size scenario. The point of recovery is around $N = 1000$ samples, where the correct estimation of the support becomes more than 50%. This is followed by a slow, but much steeper convergence to the performance of the SMB-ORACLE than the OE method. Note that this performance loss, is mainly due to the inaccurate estimation of the pre-filters and the small sample size for the BIC scheme.

VII. CONCLUSIONS

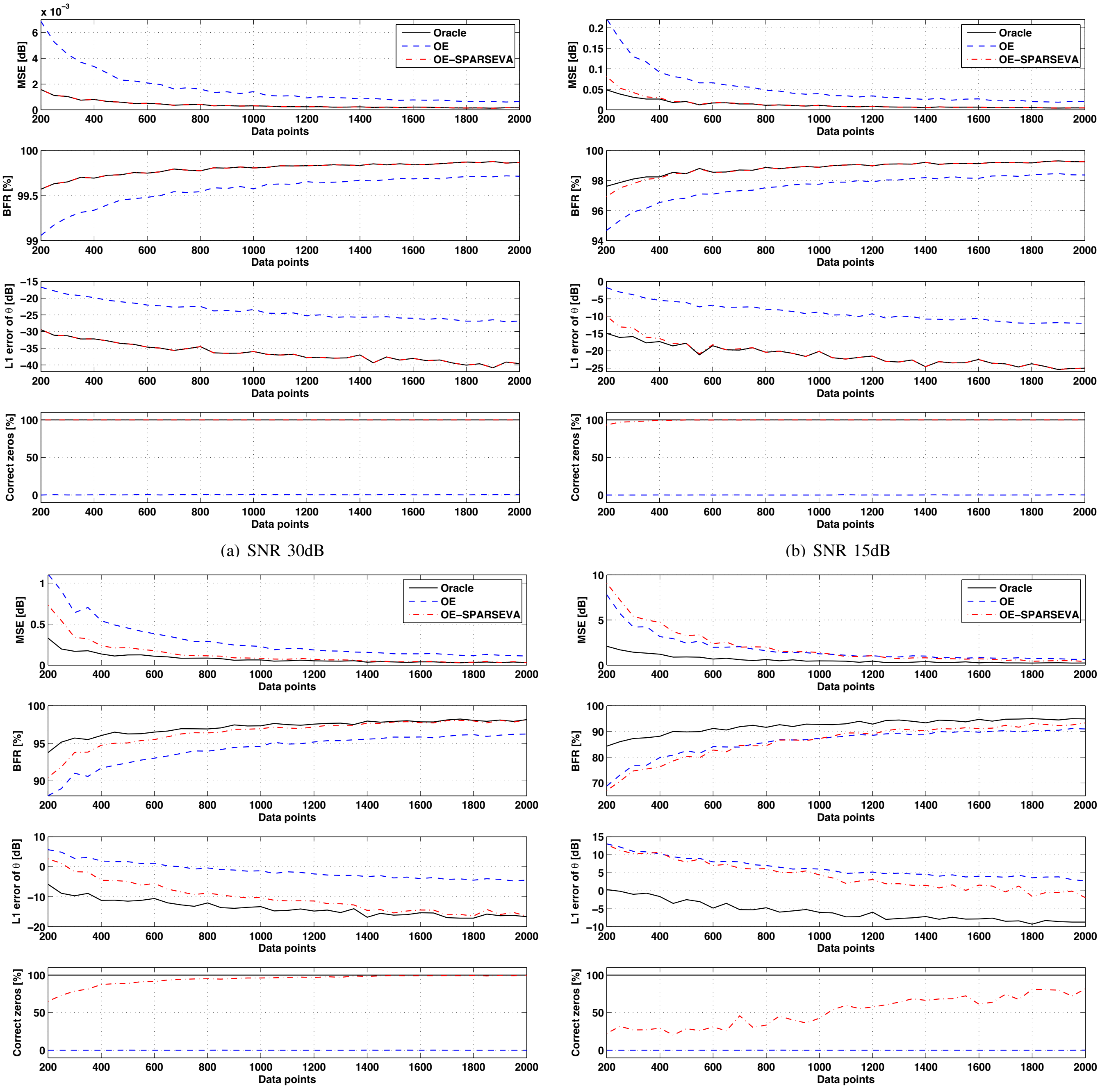
In this manuscript, we have presented two contributions to the problem of sparse estimation of rational plant structures.

The first contribution is the elimination of the need for using cross-validation to tune the regularization parameters, by proposing a new technique, called SPARSEVA, inspired by the philosophy behind Akaike's criterion. Numerical simulations have shown that the adaptive version of this method performs most favorably. On these examples, the "AIC" choice $\varepsilon_N = (1 + 2n_{\rm g}/N)V(\hat{\theta}_N^{\rm LS}, \mathcal{D}_N)$ 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. When the focus is on sparseness, the "BIC" choice $\varepsilon_N = (n_{\rm g}\log(N)/N)V(\hat{\theta}_N^{\rm LS}, \mathcal{D}_N)$ ensures this property.

As a second contribution, we have shown that by combining SPARSEVA with a high-order ARX pre-filtering based Steiglitz-McBride method, an efficient approach can be derived for the estimation of general rational LTI plant model

TABLE I
Bias and variance results of the parameter estimates by the SMB-ORACLE, OE and the OE-SPARSEVA methods in the SNR$= 7.5$dB, $N = 600$ case.

Method		a_2	a_8	b_6	b_7
θ^o		-0.1972	-0.2741	1	-8.3365
SMB-ORACLE	mean	-0.1976	-0.2714	1.0028	-8.3535
	std	0.0140	0.0169	0.1818	0.1656
OE	mean	-0.1975	-02737	0.9937	-8.3490
	std	0.0155	0.0157	0.1862	0.1652
OE-SPARSEVA	mean	-0.1972	-0.2713	0.9985	-8.3507
	std	0.0140	0.0166	0.1794	0.1676



(a) SNR 30dB

(b) SNR 15dB

(c) SNR 7.5dB

(d) SNR 0.0dB

Fig. 5. Monte Carlo simulation results with various SNR's and data lengths N.

structures in which the underlying data-generating system is represented by a sparse parameter vector. A main benefit of the method, inherited from SPARSEVA, is that the regularization parameter (or tuning quantity) is automatically chosen, not requiring cross-validation. The derived approach can be used to recover the dynamical structure of the system, *i.e.*, for model structure selection, even in case of heavy over-parametrization or colored noise settings provided that a sufficiently large data set is available. The latter has been demonstrated by an extensive simulation based Monte-Carlo study.

The theory developed in the paper is asymptotic in nature. An interesting topic for future research is to explore the small-sample / low-SNR behavior of SPARSEVA, and to consider corrected versions of AIC or BIC (as choices for ε_N) under these conditions.

Appendix A
Notation used in Appendices

The notation

$$\eta := \frac{N}{\sigma^2} V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N), \tag{33a}$$

$$\xi := [\Phi_N^\top \Phi_N]^{-1} \Phi_N^\top E_N = \hat{\theta}_N^{\mathrm{LS}} - \theta^{\mathrm{o}}, \tag{33b}$$

will be used throughout the appendices.

Appendix B

Proof of Theorem 5.1

The proof will be based on the following formulation of (12) and (13).

Lemma II.1: It holds that

$$\eta/N \to 1 \text{ in probability as } N \to \infty;$$
$$\sqrt{N}\xi \in \text{As}\,\mathcal{N}(0, \sigma^2 M), \text{ where } M \succ 0.$$

Furthermore, problems (12) and (13) can be rewritten as

$$\min_{\theta \in \mathbb{R}^{n_{\mathrm{g}}}} \quad \|\omega \odot \theta\|_1, \tag{34a}$$

$$\text{s.t.} \quad \sigma^2 \varepsilon_N \eta \geq \|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\Gamma_N}^2, \tag{34b}$$

where $\Gamma_N := \Phi_N^\top \Phi_N$, $\omega = [1 \ \cdots \ 1]^\top \in \mathbb{R}^{n_{\mathrm{g}}}$ for (12) and $\omega = w_N$ for (13).

Proof: See Appendix G. ■

To simplify the notation, define

$$\Omega := \{\theta \in \mathbb{R}^{n_{\mathrm{g}}} \mid \|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\Gamma_N} \leq \sqrt{\sigma^2 \varepsilon_N \eta}\}, \tag{35}$$

as the constraint set of (34).

Lemma II.2 (Optimality achieved on the boundary): The optimum of (34), when the elements of w are strictly positive, is achieved at $\theta = 0$ if $0 \in \Omega$, otherwise it is achieved on the boundary of Ω, and no nonzero interior point of Ω can be an optimum point of (34).

Proof: See Appendix G. ■

Since $\eta/N \xrightarrow{p} 1$ and since [7]

$$\frac{1}{N}\|z\|_{\Gamma_N}^2 = \frac{1}{N} z^\top \Gamma_N z \geq \{\lambda_{\min}(\Gamma) + o_p(1)\}\|z\|_2^2,$$

for every $z \in \mathbb{R}^{n_{\mathrm{g}}}$, (34) gives that

$$\begin{aligned}
\|\hat{\theta}_N - \theta^{\mathrm{o}}\|_2 &\leq \|\hat{\theta}_N - \hat{\theta}_N^{\mathrm{LS}}\|_2 + \|\hat{\theta}_N^{\mathrm{LS}} - \theta^{\mathrm{o}}\|_2 \\
&\leq \frac{\|\hat{\theta}_N - \hat{\theta}_N^{\mathrm{LS}}\|_{\Gamma_N}}{\sqrt{N}\big(\sqrt{\lambda_{\min}(\Gamma)} + o_p(1)\big)} + \|\hat{\theta}_N^{\mathrm{LS}} - \theta^{\mathrm{o}}\|_2 \\
&\leq \sqrt{\frac{\sigma^2 \varepsilon_N \eta}{N}}[\lambda_{\min}^{-1/2}(\Gamma) + o_p(1)] + \|\xi\|_2 \\
&= \sigma \lambda_{\min}^{-1/2}(\Gamma)\sqrt{\varepsilon_N} + o_p(1) + O_p(N^{-1/2}),
\end{aligned}$$

which implies that $\hat{\theta}_N \xrightarrow{p} \theta^{\mathrm{o}}$ if $\varepsilon_N \to 0$. Conversely, assume that $\liminf_{N\to\infty} \varepsilon_N = \delta > 0$, *i.e.*, that there is a subsequence $\{N_k\}_{k\in\mathbb{N}} \in \mathbb{N}$ such that for all $k \in \mathbb{N}$, $\varepsilon_{N_k} > \delta/2$ (say). Assume without loss of generality that N_1 is large enough so that, with probability $1 - \delta$ ($\delta > 0$), $|\|N_k^{-1}\Gamma_{N_k}\| - \|\Gamma\|| < \|\Gamma\|/2$ for all $k \in \mathbb{N}$; denote this event by S_*. Consider the neighborhood $U := \{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : \ \|\theta - \theta^{\mathrm{o}}\|_2 < \sigma\sqrt{(\delta/12)\|\Gamma\|^{-1}}\}$ of θ^{o}. Then, since for all $k \in \mathbb{N}$ and $x \in \mathbb{R}^{n_{\mathrm{g}}}$, under S_*, it holds that

$$\frac{1}{N_k}\|x\|_{\Gamma_{N_k}}^2 = \frac{1}{N_k} x^\top \Gamma_{N_k} x \leq \|N_k^{-1}\Gamma_{N_k}\|\|x\|_2^2 \leq \frac{3}{2}\|\Gamma\|\|x\|_2^2$$

and $\|\theta - (\theta^{\mathrm{o}} + \xi)\|_2 \leq \|\theta - \theta^{\mathrm{o}}\|_2 + \|\xi\|_2$, it follows that under S_*:

$$\begin{aligned}
\Omega &\supseteq \left\{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : \ \sqrt{\frac{3}{2}\|\Gamma\|}\|\theta - (\theta^{\mathrm{o}} + \xi)\|_2 \leq \sqrt{\frac{\sigma^2 \varepsilon_{N_k}\eta}{N_k}}\right\} \\
&= \left\{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : \ \|\theta - (\theta^{\mathrm{o}} + \xi)\|_2 \leq \sqrt{\frac{2\sigma^2}{3\|\Gamma\|}}\sqrt{\varepsilon_{N_k}}\sqrt{\frac{\eta}{N_k}}\right\} \\
&\supseteq \left\{\theta \in \mathbb{R}^{n_{\mathrm{g}}} : \ \|\xi\|_2 \leq \sqrt{\frac{\sigma^2\delta}{3\|\Gamma\|}}\sqrt{\frac{\eta}{N_k}} - \|\theta - \theta^{\mathrm{o}}\|_2\right\}.
\end{aligned}$$

This implies that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
P\{U \subset \Omega, S_*\} &\geq P\left\{\|\xi\|_2 < \sqrt{\frac{\sigma^2\delta}{3\|\Gamma\|}}\sqrt{\frac{\eta}{N_k}} - \sqrt{\frac{\sigma^2\delta}{12\|\Gamma\|}}, S_*\right\} \\
&= P\left\{\|\xi\|_2 < \sqrt{\frac{\sigma^2\delta}{12\|\Gamma\|}} + o_p(1), S_*\right\} \\
&\to P\{S_*\} = 1 - \delta,
\end{aligned}$$

because $\sqrt{N}\xi \in \text{As}\,\mathcal{N}(0, \sigma^2 M)$, *i.e.*, $\xi = o_p(1)$. Since δ was arbitrary, this shows that $\limsup_{N\to\infty} P\{U \subset \Omega\} = 1$, that is, $\theta = \theta^{\mathrm{o}}$ is an interior point of the constraint set of (34) with non-negligible probability for $N = N_k$ as $k \to \infty$. On the other hand, by Lemma II.2 the optimum of (34) is achieved on the boundary of Ω (or at $\theta = 0 \neq \theta^{\mathrm{o}}$) [8], which means that

$$\liminf_{N\to\infty} P\left\{\|\hat{\theta}_N^{(\mathrm{A})} - \theta^{\mathrm{o}}\|_2 > \min\left\{\sigma\sqrt{\frac{\delta}{12\|\Gamma\|}}, \|\theta^{\mathrm{o}}\|_2\right\}\right\} > 0,$$

i.e., (A-)SPARSEVA is not consistent in probability if $\varepsilon_N \nrightarrow 0$.

Proof of Corollary 5.1

The corollary follows from the fact (*c.f.* Lemma II.2) that the optimum of (34) is achieved on the boundary of the constraint set Ω (or at $\theta = 0 \neq \theta^{\mathrm{o}}$), whose size has order ε_N, larger than $N^{-1/2}$.

[7] The condition $\lim_{N\to\infty} N^{-1}\Gamma_N = \Gamma$ denotes element-wise convergence of $N^{-1}\Gamma_N$ to Γ. Since the eigenvalues of a matrix are continuous functions of its elements [34, Appendix D], such condition implies the convergence of the eigenvalues of $N^{-1}\Gamma_N$ to the eigenvalues of Γ (appropriately sorted).

[8] For A-SPARSEVA, since the elements of w_N are zero with probability 0, we can restrict ourselves to the event where they are strictly positive.

Appendix C
Proof of Theorem 5.2

Without loss of generality, let us assume that $\mathcal{T}_{\rm o} = \{1, \dots, n_1\}$, with $n_2 := n_{\rm g} - n_1 > 0$. Note that such a condition can be satisfied by reordering the columns of $\varPhi$ in (6). Then, the following result holds:

Lemma III.1 (Conditions for sparseness): Under the stated assumptions for (34), and $\theta^{\rm o} \neq 0$, the optimal solution of (34) (with $\omega = w_N$), $\hat{\theta}_N^{\rm A} = [(\hat{\theta}_N^{{\rm A},1})^\top \ (\hat{\theta}_N^{{\rm A},2})^\top]^\top$, with $\hat{\theta}_N^{{\rm A},i} \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, satisfies ${\rm Supp}(\hat{\theta}_N^{{\rm A},2}) = \mathbb{I}_1^{n_1}$ and ${\rm Supp}(\hat{\theta}_N^{{\rm A},2}) = \emptyset$, *i.e.*, recovery of the true support $\mathcal{T}_{\rm o}$ holds, if the following conditions hold:

- $\varGamma_N > 0$,
- $\sqrt{\sigma^2 \varepsilon_N \eta} < 0.5\sqrt{\lambda_{\min}(\varGamma_N)}\min\{|[\theta^{\rm o}]_i| : \ i \in \mathbb{I}_1^{n_1}\}$,
- $|\xi_i| < |[\theta^{\rm o}]_i|, \quad$ for all $i \in \mathbb{I}_1^{n_1}$,
- $\sqrt{n}{\rm Cond}(\varGamma_N)\|\xi_{\rm a}^{(2)}\|_\infty +$
$$\sqrt{\frac{\sigma^2\varepsilon_N\eta}{\lambda_{\min}(\varGamma_N^{-1})}}\frac{\{1 + {\rm Cond}(\varGamma_N)\}\|\varGamma_N^{-1}\|\|w_N^{(1)}\|_2}{\sqrt{\|w_N^{(1)}\|_2^2 + \min_{i\in\mathbb{I}_1^{n_2}}|[w_N^{(2)}]_i|^2}}$$
$$\leq \sqrt{\sigma^2\varepsilon_N\eta}\frac{\lambda_{\min}(\varGamma_N^{-1})}{\sqrt{\|\varGamma_N^{-1}\|}}\frac{\min_{i\in\mathbb{I}_1^{n_2}}|[w_N^{(2)}]_i|^2}{\sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}},$$

where $w_N =: [(w_N^{(1)})^\top \ (w_N^{(2)})^\top]^\top$, according to the partition of $\hat{\theta}_N^{\rm A}$ and $\xi_{\rm a}^{(2)}$ corresponds to those ξ_i which are associated with the parameters of A in $\theta^{{\rm o},2}$.

Proof: See Appendix G. ■

Let us first assume that $N\varepsilon_N \to \infty$. To establish the sparseness of A-SPARSEVA, we just need to show that the conditions of Lemma III.1 hold with probability tending to 1 as $N \to \infty$. In particular, these conditions can be written as

- $\varGamma + o_p(1) > 0$,
- $\sigma\sqrt{\varepsilon_N} + o_p(1) < 0.5\sqrt{\lambda_{\min}(\varGamma)}\min\limits_{i\in\mathbb{I}_1^{n_1}}|[\theta^{\rm o}]_i| + o_p(1)$,
- $O_p(N^{-1/2}) < |[\theta^{\rm o}]_i|$, for all $i \in \mathbb{I}_1^{n_1}$,
- $O_p(N^{-1/2}) + \sigma\sqrt{\lambda_{\max}(\varGamma)}\{1 + {\rm Cond}(\varGamma)\}\lambda_{\max}(\varGamma^{-1})\cdot$
- $\sqrt{\varepsilon_N}O_p(N^{-\gamma/2}) \leq \sigma\sqrt{\varepsilon_N}\dfrac{\lambda_{\min}(\varGamma^{-1})}{\sqrt{\lambda_{\max}(\varGamma^{-1})}}(1 + o_p(1))$.

Since $\varepsilon_N \to 0$, $N\varepsilon_N \to \infty$ and $\gamma > 0$, all these conditions hold (separately) with probability tending to 1 as $N \to \infty$. By Boole's inequality [35], the probability that all of them hold simultaneously tends to 1 as $N \to \infty$. Hence, by Lemma III.1, A-SPARSEVA has the sparseness property.

Let us assume now that $N\varepsilon_N \to \infty$ does not hold. Problem (34) can be written as

$$\begin{aligned}\min_{\theta_1,\theta_2} \quad & \|\omega^{(1)} \odot \theta_1\|_1 + \|\omega^{(2)} \odot \theta_2\|_1 \\ \text{s.t.} \quad & \sigma^2\varepsilon_N\eta \geq \begin{bmatrix}\theta_1 - (\theta^{{\rm o},1} + \xi^{(1)}) \\ \theta_2 - \xi^{(2)}\end{bmatrix}^\top \varGamma_N \begin{bmatrix}\theta_1 - (\theta^{{\rm o},1} + \xi^{(1)}) \\ \theta_2 - \xi^{(2)}\end{bmatrix}\end{aligned} \tag{36}$$

By Theorem 5.1, $\|\hat{\theta}_N^{\rm A} - \theta^{\rm o}\|_2 = O_p(N^{-1/2})$, which implies that $\|\hat{\theta}_N^{{\rm A},1} - (\theta^{{\rm o},1} + \xi^{(1)})\|_2^2 = O_p(N^{-1})$. Letting $\hat{\theta}_N^{{\rm A},1}$ fixed, it follows from (36) that $\hat{\theta}_N^{{\rm A},2} = 0$ if and only if $\hat{\theta}_N^{\rm A} = [(\hat{\theta}_N^{{\rm A},1})^\top \ 0]^\top$ satisfies the constraint in (36), *i.e.*, if

$$\begin{aligned}&\sigma^2\varepsilon_N N + o_p(N) \\ &\quad\geq \begin{bmatrix}\hat{\theta}_N^{{\rm A},1} - (\theta^{{\rm o},1} + \xi^{(1)}) \\ -\xi^{(2)}\end{bmatrix}^\top \varGamma_N \begin{bmatrix}\hat{\theta}_N^{{\rm A},1} - (\theta^{{\rm o},1} + \xi^{(1)}) \\ -\xi^{(2)}\end{bmatrix} \\ &\quad\geq \big(\lambda_{\min}(\varGamma) + o_p(1)\big)\big(N\|\hat{\theta}_N^{{\rm A},1} - (\theta^{{\rm o},1} + \xi^{(1)})\|_2^2 + N\|\xi^{(2)}\|_2^2\big) \\ &\quad= \sigma^2\lambda_{\min}(\varGamma)N\|\xi^{(2)}\|_2^2 + O_p(1).\end{aligned} \tag{37}$$

As $\sqrt{N}\xi^{(2)} \in {\rm As}\,\mathcal{N}(0, \sigma^2 M_2)$ for some $M_2 > 0$, $N\|\xi^{(2)}\|_2^2/\sigma^2$ has asymptotically a distribution with unbounded support, hence the probability that (37) holds does not tend to 1 as $N \to \infty$. Hence, in this case A-SPARSEVA does not have the sparseness property (only if $N\varepsilon_N \to \infty$).

Appendix D
Proof of Theorem 5.3

Denote by $\hat{\theta}_N^{\rm oracle}$ the least squares estimate, which is obtained with the exact knowledge of the true support $\mathcal{T}_{\rm o}$, *i.e.*, with (9) using $\varPhi_{N,\mathcal{T}_{\rm o}}$. Furthermore, let M be the asymptotic information matrix of θ assuming knowledge of $\mathcal{T}_{\rm o}$ and S_N be the event that ${\rm Supp}(\hat{\theta}_N^{\rm A}) = \mathcal{T}_{\rm o}$. The complement of S_N is denoted by $\bar{S}_N$. This gives that, for every $x \in \mathbb{R}^{n_{\rm g}}$,

$$\begin{aligned}&P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\} \\ &\quad= P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x | S_N\}P\{S_N\} + \\ &\qquad P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x, \bar{S}_N\} \\ &\quad= P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\}P\{S_N\} + \\ &\qquad P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\}P\{\bar{S}_N\}.\end{aligned}$$

where $\leq$ is taken component-wisely, and $M^{1/2}$ denotes the positive definite square root of M. Therefore,

$$\begin{aligned}&|P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\} \\ &\quad - P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\}| \\ &= |P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\}(P\{S_N\} - 1) \\ &\quad + P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\}P\{\bar{S}_N\}| \\ &\leq [P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\} + 1]P\{\bar{S}_N\} \\ &\leq 2P\{\bar{S}_N\}.\end{aligned} \tag{38}$$

Hence, if $F(x)$ denotes the cumulative standard normal distribution function, then (38) implies that

$$\begin{aligned}&\lim_{N\to\infty}\Big|P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\} - F(x)\Big| \\ &\leq \lim_{N\to\infty}|P\{[NM]^{1/2}(\hat{\theta}_N^{\rm A-RE} - \theta^{\rm o}) \leq x\} \\ &\qquad\qquad - P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\}| \qquad (39) \\ &\quad + \lim_{N\to\infty}\Big|P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\} - F(x)\Big| \\ &\leq 2\lim_{N\to\infty}P\{\bar{S}_N\} + \\ &\quad \lim_{N\to\infty}\Big|P\{[NM]^{1/2}(\hat{\theta}_N^{\rm oracle} - \theta^{\rm o}) \leq x\} - F(x)\Big| = 0,\end{aligned}$$

since $\hat{\theta}_N^{\rm A}$ has the sparseness property (*i.e.*, $\lim_{N\to\infty}P\{\bar{S}_N\} = 0$), and $\hat{\theta}_N^{\rm oracle}$ is asymptotically efficient and normal. Equation (39) shows that $\hat{\theta}_N^{\rm A-RE}$ has the Oracle property.

APPENDIX E
PROOF OF THEOREM 5.4

First notice that $\inf_{\hat{\delta}\in\mathbb{R}^{n_\mathrm{g}}} \sup_{\theta^\mathrm{o}\in\mathbb{R}^{n_\mathrm{g}}} R(\hat{\delta},\theta^\mathrm{o}) \asymp_p N^{-1}$. To see this, note that the estimator $\hat{\theta}_N^\mathrm{LS}$ is minimax optimal because it coincides with the maximum likelihood estimator of θ^o (which is known to be minimax, see *e.g.* [36, Section 5.3.2]), and for this estimator, the worst-case risk decays asymptotically as N^{-1}.

Consider now the case $\varepsilon_N = O_p(N^{-1})$. Then, by Theorem 5.1, $\sup_{\theta^\mathrm{o}\in\mathbb{R}^{n_\mathrm{g}}} R(\hat{\theta}_N^\mathrm{A},\theta^\mathrm{o}) = O_p(N^{-1})$, which shows that A-SPARSEVA is minimax rate optimal. To show that A-SPARSEVA-RE is also minimax rate optimal, denote by $\mathbb{L}$ the subspace of $\mathbb{R}^{n_\mathrm{g}}$ consisting of all points $\theta \in \mathbb{R}^{n_\mathrm{g}}$ such that $\theta_{\bar{\mathcal{T}}_\mathrm{A}} = 0$ with $\mathcal{T}_\mathrm{A} = \mathrm{Supp}(\hat{\theta}_N^\mathrm{A})$; in other words, $\mathbb{L}$ is the set of parameter values with the same support as A-SPARSEVA. By definition, for all N sufficiently large, $\mathbb{L} \cap \varOmega \neq \emptyset$ with high probability, because $\hat{\theta}_N^\mathrm{A} \in \mathbb{L} \cap \varOmega$ with high probability. This implies that, for N sufficiently large, $\hat{\theta}_N^{\mathrm{A-RE}} \in \varOmega$ with high probability, since otherwise $V(\hat{\theta}_N^{\mathrm{A-RE}},\mathcal{D}_N) > V(\hat{\theta}_N^\mathrm{A},\mathcal{D}_N)$, contradicting the optimality of $V(\hat{\theta}_N^{\mathrm{A-RE}},\mathcal{D}_N)$. This observation implies, by following a similar argument as in the proof of Theorem 5.1, that $\sup_{\theta^\mathrm{o}\in\mathbb{R}^{n_\mathrm{g}}} R(\hat{\theta}_N^{\mathrm{A-RE}},\theta^\mathrm{o}) = O_p(N^{-1})$.

The necessity of the condition $\varepsilon_N = O_p(N^{-1})$ for minimax rate optimality follows from [29, Theorem 2.1] (see also [30, Theorem 1]). To apply this theorem, restrict ε_N to a subsequence $\{N_t\}_{t\in\mathbb{N}}$ such that $N_t\varepsilon_{N_t} \to \infty$ as $t \to \infty$. For this subsequence, according to Theorem 5.2, both A-SPARSEVA and A-SPARSEVA-RE have the sparseness property, which implies, by [29, Theorem 2.1], that $N_t \sup_{\theta^\mathrm{o}\in\mathbb{R}^{n_\mathrm{g}}} R(\hat{\theta}_{N_t}^\mathrm{A},\theta^\mathrm{o}) \to \infty$. Hence, A-SPARSEVA and A-SPARSEVA-RE are not minimax rate optimal, unless $\varepsilon_N = O_p(N^{-1})$.

APPENDIX F
PROOF OF THEOREM 5.5

The three properties follow from Theorems 5.1, 5.2 and 5.3 (with the asymptotic efficiency of the modified Steiglitz-McBride method), respectively, if the assumptions of Section IV-A1 hold. Therefore, we need to show that such assumptions are valid.

The third assumption in Section III-A1 follows directly from the asymptotic efficiency of the modified Steiglitz-McBride method.

To verify the first two assumptions, notice that the Steiglitz-McBride iterations (steps 4-7 of Algorithm 2) deliver a polynomial $\hat{A}^{(k+1)}(q)$ which is an asymptotically efficient estimate of $A_\mathrm{o}(q)$, *i.e.*, $\hat{A}^{(k+1)}(q) = A_\mathrm{o}(q) + O_p(N^{-1/2})$. In addition, the data satisfies asymptotically an ARX structure of the form

$$A_\mathrm{o}(q)y_t^{\mathrm{o,F}} = B_\mathrm{o}(q)u_t^{\mathrm{o,F}} + v_t,$$

where $v_t = H_N(q)e_t$, with $\{e_t\}$ a Gaussian white noise sequence of variance σ^2 and $\{H_N(q)\}$ a sequence of filters such that $\sup_{|z|=1} |H_N(z) - 1| = o_p(1)$ [37, Theorem 3.1]. Therefore, the application of the filter $1/\hat{A}^{(k+1)}(q)$ yields data $\{u_t^\mathrm{F}, y_t^\mathrm{F}\}$ such that

$$\begin{aligned} u_t^{\mathrm{F}(k)} &= \tilde{u}_t^{\mathrm{o,F}} + o_p(1) \\ y_t^{\mathrm{F}(k)} &= \tilde{y}_t^{\mathrm{o,F}} + o_p(1), \end{aligned}$$

where $\{\tilde{u}_t^{\mathrm{o,F}}, \tilde{y}_t^{\mathrm{o,F}}\}$ are such that

$$A_\mathrm{o}(q)\tilde{y}_t^{\mathrm{o,F}} = B_\mathrm{o}(q)\tilde{u}_t^{\mathrm{o,F}} + e_t.$$

The regressor matrix fed to A-SPARSEVA-RE then satisfies

$$\varPhi_N^{(k)} = \varPhi_N^\mathrm{o} + o_p(1),$$

where $\varPhi_N^0$ is the regressor matrix obtained from $\{\tilde{u}_t^{\mathrm{o,F}}, \tilde{y}_t^{\mathrm{o,F}}\}$. Therefore,

$$\frac{1}{N}(\varPhi_N^{(k)})^\top \varPhi_N^{(k)} = \frac{1}{N}(\varPhi_N^\mathrm{o})^\top \varPhi_N^\mathrm{o} + o_p(1),$$

due to the law of large numbers (see *e.g.* [2, Theorem 2.B.1]). Appealing again to [2, Theorem 2.B.1], we obtain

$$\frac{1}{N}(\varPhi_N^{(k)})^\top \varPhi_N^{(k)} \xrightarrow{p} \varGamma > 0,$$

for some $\varGamma > 0$. This verifies the second assumption of Section IV-A1. Finally, notice that

$$\begin{aligned} V(\theta,\mathcal{D}_N) &= \frac{1}{N}\|Y_N^\mathrm{F} - \varPhi_N^{(k)}\theta\|_2^2 \\ &= \frac{1}{N}\|\theta - \big((\varPhi_N^{(k)})^\top \varPhi_N\big)^{-1}(\varPhi_N^{(k)})^\top Y_N^\mathrm{F}\|_2^2 \\ &\quad + \frac{1}{N}(Y_N^\mathrm{F})^\top[I - \varPhi_N^{(k)}\big((\varPhi_N^{(k)})^\top \varPhi_N^{(k)}\big)^{-1}(\varPhi_N^{(k)})^\top]Y_N^\mathrm{F}, \end{aligned}$$

where $Y_N^\mathrm{F} = [y_1^\mathrm{F} \ \ \cdots \ \ y_N^\mathrm{F}]^\top$, hence

$$V(\hat{\theta}_N^\mathrm{LS},\mathcal{D}_N) = \frac{1}{N}(Y_N^\mathrm{F})^\top[I - \varPhi_N^{(k)}\big((\varPhi_N^{(k)})^\top \varPhi_N^{(k)}\big)^{-1}(\varPhi_N^{(k)})^\top]Y_N^\mathrm{F}.$$

Now,

$$\begin{aligned} &\varPhi_N^{(k)}\big((\varPhi_N^{(k)})^\top \varPhi_N^{(k)}\big)^{-1}(\varPhi_N^{(k)})^\top \\ &= \frac{1}{N}\varPhi_N^{(k)}\left(\frac{1}{N}(\varPhi_N^{(k)})^\top \varPhi_N^{(k)}\right)^{-1}(\varPhi_N^{(k)})^\top \\ &= \frac{1}{N}\{\varPhi_N^\mathrm{o} + o_p(1)\}\left\{\frac{1}{N}(\varPhi_N^\mathrm{o})^\top \varPhi_N^\mathrm{o} + o_p(1)\right\}^{-1}\{\varPhi_N^\mathrm{o} + o_p(1)\}^\top \\ &= \frac{1}{N}\{\varPhi_N^\mathrm{o} + o_p(1)\}\left\{\left(\frac{1}{N}(\varPhi_N^\mathrm{o})^\top \varPhi_N^\mathrm{o}\right)^{-1} + o_p(1)\right\}\{\varPhi_N^\mathrm{o} + o_p(1)\}^\top \\ &= \varPhi_N^\mathrm{o}\big((\varPhi_N^\mathrm{o})^\top \varPhi_N^\mathrm{o}\big)^{-1}\varPhi_N^\mathrm{o})^\top + o_p(1), \end{aligned}$$

where we have used [2, Theorem 2.B.1] in the last step. Therefore,

$$\begin{aligned} V(\hat{\theta}_N^\mathrm{LS},\mathcal{D}_N) &= \frac{1}{N}(Y_N^F)^\top[I - \varPhi_N^{(k)}\big((\varPhi_N^{(k)})^\top \varPhi_N^{(k)}\big)^{-1}(\varPhi_N^{(k)})^\top]Y_N^F \\ &= \frac{1}{N}((Y_N^\mathrm{o})^\top + o_p(1)) \\ &\cdot [I - \varPhi_N^\mathrm{o}\big((\varPhi_N^\mathrm{o})^\top \varPhi_N^\mathrm{o}\big)^{-1}(\varPhi_N^\mathrm{o})^\top + o_p(1)] \cdot (Y_N^\mathrm{o} + o_p(1)) \\ &= \sigma^2 + o_p(1), \end{aligned}$$

using [2, Theorem 2.B.1] again. This verifies the first assumption of Section IV-A1, which concludes the proof.

Appendix G
Proof of auxiliary results

Proof of Lemma II.1

The first two assertions follow from the assumptions in Section IV-A1. Furthermore, expanding (8) and using that $\big(I - \varPhi_N(\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top\big)$ is idempotent gives

$$N \cdot V(\theta, \mathcal{D}_N) = \|\theta - (\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top Y_N\|_{\varGamma_N}^2 + \\ \|\big(I - \varPhi_N(\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top\big)E_N\|_2^2, \quad (40)$$

For $\hat{\theta}_N^{\mathrm{LS}} = (\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top Y_N$, (40) reveals that

$$\frac{1}{\sigma^2} V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N) = \frac{1}{N}\left\| \frac{1}{\sigma}(I - \varPhi_N(\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top)E_N \right\|_2^2.$$

These two observations imply that

$$(1 + \varepsilon_N) V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N) \geq V(\theta, \mathcal{D}_N),$$

or, equivalently,

$$\varepsilon_N V(\hat{\theta}_N^{\mathrm{LS}}, \mathcal{D}_N) \geq \big\|\theta - \big(\theta^{\mathrm{o}} + (\varPhi_N^\top \varPhi_N)^{-1}\varPhi_N^\top E_N\big)\big\|_{\varGamma_N}^2.$$

Proof of Lemma II.2

If $0 \in \varOmega$, then it trivially follows that $\theta = 0$ is the unique optimum of (34). Now let us assume for the rest of the proof that $0 \notin \varOmega$. Notice that $\|\omega \odot \cdot\|_1$ is a norm in $\mathbb{R}^{n_{\mathrm{g}}}$. Since $\varOmega$ is a closed set in the topology of $\|\cdot\|_{\varGamma_N}$, it is also closed in the topology of $\|\omega \odot \cdot\|_1$ (since all norms in $\mathbb{R}^{n_{\mathrm{g}}}$ are topologically equivalent [38, Problem 15.7]). Hence, $\mathrm{dist}(\varOmega, 0) = \inf\{\|w \odot \theta\|_1 : \theta \in \varOmega\} =: \delta > 0$, and there is an element $\theta^* \in \varOmega$ such that $\|w \odot \theta^*\|_1 = \delta$. Such θ^* is an optimum of (34).

Let us assume that an interior point of $\varOmega$, say $\bar{\theta}$, achieves $\|\omega \odot \bar{\theta}\|_1 = \delta$, and consider a neighborhood $U := \{\theta \in \varOmega : \|\omega \odot (\theta - \bar{\theta})\|_1 < \lambda\} \subset \varOmega$, where $\lambda < \delta$. Then the point $\tilde{\theta} = \big((\|\omega \odot \bar{\theta}\|_1 - \lambda/2)/\|\omega \odot \bar{\theta}\|_1\big)\bar{\theta}$ satisfies

$$\begin{aligned}\|\omega \odot (\tilde{\theta} - \bar{\theta})\|_1 &= \left\|\omega \odot \left[\left(\frac{\|\omega \odot \bar{\theta}\|_1 - \lambda/2}{\|\omega \odot \bar{\theta}\|_1}\right)\bar{\theta} - \bar{\theta}\right]\right\|_1 \\ &= \left\|\frac{\|\omega \odot \bar{\theta}\|_1 - \lambda/2}{\|\omega \odot \bar{\theta}\|_1} - 1\right\| \|\omega \odot \bar{\theta}\|_1 = \frac{\lambda}{2},\end{aligned}$$

hence $\tilde{\theta} \in U \subset \varOmega$, but

$$\begin{aligned}\|\omega \odot \tilde{\theta}\|_1 &= \left\|\omega \odot \left[\left(\frac{\|w \odot \bar{\theta}\|_1 - \lambda/2}{\|\omega \odot \bar{\theta}\|_1}\right)\bar{\theta}\right]\right\|_1, \\ &= |\delta - \lambda/2| < \delta.\end{aligned}$$

Based on the above relation, $\tilde{\theta}$ achieves a lower cost than $\bar{\theta}$. This contradiction implies that no interior point of $\varOmega$ can be optimal.

Proof of Lemma III.1

First notice that due to the second and third assumptions of the lemma, $\|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\varGamma_N}^2 \leq \sigma^2 \varepsilon_N \eta$ can hold only if $\theta = [\theta^{(1)}\ \theta^{(2)}]^\top$ with $\theta^{(i)} \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_k^{(1)} \neq 0$ for every $k \in \mathbb{I}_1^{n_1}$. Otherwise, if $[\theta^{(1)}]_k = 0$ for some $k \in \mathbb{I}_1^{n_1}$,

$$\begin{aligned}|[\theta^{\mathrm{o}}]_k| = |[\theta^{(1)}]_k - [\theta^{\mathrm{o}}]_k| &\leq \sqrt{\sum_{i=1}^{n}[\theta_i - ([\theta^{\mathrm{o}}]_i + \xi_i)]^2} \\ &= \|\theta - (\theta^{\mathrm{o}} + \xi)\|_2 \leq \frac{1}{\sqrt{\lambda_{\min}(\varGamma_N)}}\|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\varGamma_N} \\ &\leq 0.5 \min_{i \in \mathbb{I}_1^{n_1}}(|[\theta_1^{\mathrm{o},1}]_i|),\end{aligned}$$

which is a contradiction. Now, let us further partition $\theta^{(2)}$ as $\theta^{(2)} =: [(\theta_{\mathrm{a}}^{(2)})^\top\ (\theta_{\mathrm{b}}^{(2)})^\top]^\top$, where θ_{a}^2 corresponds to those $[\theta^{(2)}]_i$ which are associated with the parameters of A, and partition w_N and ξ accordingly. Note that these parameters are exactly the zero parameters allocated at A. $\theta_{\mathrm{a}}^{(2)}$ is defined respectively. Our goal is to show that this partition leads to a contradiction if $n_{\mathrm{a}} > 0$.

By the assumptions of the lemma, $0 \notin \varOmega$, hence by Lemma II.2, the optimum lies in the boundary of $\varOmega$. Therefore, the optimality conditions for problem (34), omitting the complementary conditions, are [39, Section 28]

$$\begin{aligned}&0 \in \partial_\theta(\|\omega \odot \theta\|_1 + (\mu/2)[\|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\varGamma_N}^2 - \sigma^2 \varepsilon_N \eta]), \\ &\mu \geq 0, \quad \|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\varGamma_N}^2 = \sigma^2 \varepsilon_N \eta, \end{aligned} \quad (41)$$

for some μ, where $\partial_\theta f(\theta)$ denotes the subdifferential of a function with respect to θ. Note that the regularity conditions for (41) to be necessary and sufficient hold, since the constraint set contains an interior point, e.g., $\theta = \theta^{\mathrm{o}} + \xi$. After some algebra, using facts such as $\partial\|x\|_1 = \mathrm{Sgn}(x)$, and the partition of θ, θ^o and w, we can rewrite (41) as

$$\begin{aligned}&\begin{bmatrix}\theta^{(1)} - (\theta^{\mathrm{o},1} + \xi^{(1)}) \\ \theta_{\mathrm{a}}^{(2)} - \xi_{\mathrm{a}}^{(2)} \\ -\xi_{\mathrm{b}}^{(2)}\end{bmatrix} = -\frac{1}{\mu}\varGamma_N^{-1}\begin{bmatrix} w_N^{(1)} \odot \mathrm{Sgn}(\theta^{(1)}) \\ w_{N,\mathrm{a}}^{(2)} \odot \mathrm{Sgn}(\theta_{\mathrm{a}}^{(2)}) \\ w_{N,\mathrm{b}}^{(2)} \odot \mathrm{Sgn}(0)\end{bmatrix} \\ &\mu \geq 0, \qquad \|\theta - (\theta^{\mathrm{o}} + \xi)\|_{\varGamma_N}^2 = \sigma^2 \varepsilon_N \eta. \end{aligned} \quad (42)$$

We can partition $\varGamma_N^{-1}$ according to the partition of θ as

$$\varGamma_N^{-1} =: \begin{bmatrix} M_{11} & M_{1,\mathrm{a}} & M_{1,\mathrm{b}} \\ M_{\mathrm{a},1} & M_{\mathrm{a},\mathrm{a}} & M_{\mathrm{a},\mathrm{b}} \\ M_{\mathrm{b},1} & M_{\mathrm{b},\mathrm{a}} & M_{\mathrm{b},\mathrm{b}} \end{bmatrix},$$

which, together with (42), gives

$$\begin{aligned}\theta_{\mathrm{a}}^{(2)} = &-\frac{1}{\mu}[M_{\mathrm{a},\mathrm{a}} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}M_{\mathrm{b},\mathrm{a}}][w_{N,\mathrm{a}}^{(2)} \odot \mathrm{Sgn}(\theta_{\mathrm{a}}^{(2)})] \\ &+ \xi_{\mathrm{a}}^{(2)} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}\xi_{\mathrm{b}}^{(2)} \\ &- \frac{1}{\mu}[M_{\mathrm{a},1} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}M_{\mathrm{b},1}][w_N^{(1)} \odot \mathrm{Sgn}(\theta^{\mathrm{o},1})]\end{aligned} \quad (43)$$

Note that $M_{\mathrm{a},\mathrm{a}} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}M_{\mathrm{b},\mathrm{a}} > 0$ (since $\varGamma_N > 0$ [34, Theorem 7.7.6]). Therefore, (43) and the assumptions of the lemma imply the sparseness of $\hat{\theta}_N^{\mathrm{A}}$ by noting that

$$\begin{aligned}&0 < \sum_{i=1}^{n_{\mathrm{a}}}[w_{N,\mathrm{a}}^{(2)}]_i|[\theta_{\mathrm{a}}^{(2)}]_i| = q^\top \theta_{\mathrm{a}}^{(2)} = \quad (44) \\ &-\frac{1}{\mu}q^\top[M_{\mathrm{a},\mathrm{a}} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}M_{\mathrm{b},\mathrm{a}}]q + q^\top\Big(\xi_{\mathrm{a}}^{(2)} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}\xi_{\mathrm{b}}^{(2)} \\ &\qquad -\frac{1}{\mu}[M_{\mathrm{a},1} - M_{\mathrm{a},\mathrm{b}}M_{\mathrm{b},\mathrm{b}}^{-1}M_{\mathrm{b},1}][w_N^{(1)} \odot \mathrm{Sgn}(\theta^{\mathrm{o},1})]\Big) \\ &< 0,\end{aligned}$$

where $q = w_{N,\mathrm{a}}^{(2)} \odot \mathrm{Sgn}(\theta_{\mathrm{a}}^{(2)})$, which is a contradiction. However, to establish (44) we still need to prove that the first term of the second line of (44) dominates the second term. To this end, notice first that

$$\begin{aligned}
\mu &\geq \frac{1}{\sqrt{\sigma^2 \varepsilon_N \eta}} \sqrt{\lambda_{\min}(\mathit{\Gamma}_N^{-1})} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_{N,\mathrm{a}}^{(2)}\|_2^2} \\
\mu &\leq \frac{1}{\sqrt{\sigma^2 \varepsilon_N \eta}} \sqrt{\|\mathit{\Gamma}_N^{-1}\|} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}.
\end{aligned}$$

Based on these inequalities, we have

$$\begin{aligned}
&\frac{1}{\mu} |q^\top [M_{\mathrm{a,a}} - M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1} M_{\mathrm{b,a}}] q| \\
&\geq \frac{1}{\mu} \lambda_{\min}\{M_{\mathrm{a,a}} - M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1} M_{\mathrm{b,a}}\} \|w_{N,\mathrm{a}}^{(2)}\|_2^2 \\
&\geq \frac{\sqrt{\sigma^2 \varepsilon_N \eta} \|w_{N,\mathrm{a}}^{(2)}\|_2^2}{\sqrt{\|\mathit{\Gamma}_N^{-1}\|} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}} \frac{1}{\|(M_{\mathrm{a,a}} - M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1} M_{\mathrm{b,a}})^{-1}\|} \\
&\geq \frac{\sqrt{\sigma^2 \varepsilon_N \eta} \|w_{N,\mathrm{a}}^{(2)}\|_2^2}{\sqrt{\|\mathit{\Gamma}_N^{-1}\|} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}} \frac{1}{\left\| \begin{bmatrix} M_{\mathrm{a,a}} & M_{\mathrm{a,b}} \\ M_{\mathrm{b,a}} & M_{\mathrm{b,b}} \end{bmatrix}^{-1} \right\|} \\
&\geq \frac{\sqrt{\sigma^2 \varepsilon_N \eta} \|w_{N,\mathrm{a}}^{(2)}\|_2^2}{\sqrt{\|\mathit{\Gamma}_N^{-1}\|} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}} \lambda_{\min} \left\{ \begin{bmatrix} M_{\mathrm{a,a}} & M_{\mathrm{a,b}} \\ M_{\mathrm{b,a}} & M_{\mathrm{b,b}} \end{bmatrix} \right\} \\
&\geq \sqrt{\sigma^2 \varepsilon_N \eta} \frac{\|w_{N,\mathrm{a}}^{(2)}\|_2 \min_{i \in \mathbb{I}_1^{n_2}} |[w_N]_i|^2}{\sqrt{\|w_N^{(1)}\|_2^2 + \|w_N^{(2)}\|_2^2}} \frac{\lambda_{\min}(\mathit{\Gamma}_N^{-1})}{\sqrt{\|\mathit{\Gamma}_N^{-1}\|}},
\end{aligned}$$

(using [34, Section 0.7.3 and Theorem 4.3.15]) and

$$\begin{aligned}
&\Bigg| q^\top \Bigg(\xi_{\mathrm{a}}^{(2)} - M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1} \xi_{\mathrm{b}}^{(2)} \\
&\qquad\qquad - \frac{1}{\mu} [M_{\mathrm{a},1} - M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1} M_{\mathrm{b},1}] [w_N^{(1)} \odot \mathrm{Sgn}(\theta^{\mathrm{o},1})] \Bigg) \Bigg| \\
&< \|\xi_{\mathrm{a}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 + \|M_{\mathrm{a,b}} M_{\mathrm{b,b}}^{-1}\| \|\xi_{\mathrm{b}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 + \\
&\quad \frac{\sqrt{\sigma^2 \varepsilon_N \eta} \left\{ \|M_{\mathrm{a},1}\| + \|M_{\mathrm{a,b}}\| \|M_{\mathrm{b,b}}^{-1}\| \|M_{\mathrm{b},1}\| \right\} \|w_N^{(1)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2}{\sqrt{\lambda_{\min}(\mathit{\Gamma}_N^{-1})} \sqrt{\|w_N^{(1)}\|_2^2 + \|w_{N,\mathrm{a}}^{(2)}\|_2^2}} \\
&< \|\xi_{\mathrm{a}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 + \frac{\|M_{\mathrm{a,b}}\|}{\lambda_{\min}(M_{\mathrm{b,b}})} \|\xi_{\mathrm{b}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 \\
&\qquad + \left(\|M_{\mathrm{a},1}\| + \frac{\|M_{\mathrm{a,b}}\| \|M_{\mathrm{b},1}\|}{\lambda_{\min}(M_{\mathrm{b,b}})} \right) \frac{\sqrt{\sigma^2 \varepsilon_N \eta} \|w_N^{(1)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2}{\sqrt{\|w_N^{(1)}\|_2^2 + \|w_{N,\mathrm{a}}^{(2)}\|_2^2}} \\
&< \|\xi_{\mathrm{a}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 + \frac{\|\mathit{\Gamma}_N^{-1}\|}{\lambda_{\min}(\mathit{\Gamma}_N^{-1})} \|\xi_{\mathrm{b}}^{(2)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2 + \\
&\quad \sqrt{\frac{\sigma^2 \varepsilon_N \eta}{\lambda_{\min}(\mathit{\Gamma}_N^{-1})}} \left\{ 1 + \frac{\|\mathit{\Gamma}_N^{-1}\|}{\lambda_{\min}(\mathit{\Gamma}_N^{-1})} \right\} \frac{\|\mathit{\Gamma}_N^{-1}\| \|w_N^{(1)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2}{\sqrt{\|w_N^{(1)}\|_2^2 + \|w_{N,\mathrm{a}}^{(2)}\|_2^2}} \\
&\leq \sqrt{n_{\mathrm{g}}} \mathrm{Cond}(\mathit{\Gamma}_N) \|\xi_{\mathrm{a}}^{(2)}\|_\infty \|w_{N,\mathrm{a}}^{(2)}\|_2 + \\
&\quad \sqrt{\frac{\sigma^2 \varepsilon_N \eta}{\lambda_{\min}(\mathit{\Gamma}_N^{-1})}} \frac{1 + \mathrm{Cond}(\mathit{\Gamma}_N)\} \|\mathit{\Gamma}_N^{-1}\| \|w_N^{(1)}\|_2 \|w_{N,\mathrm{a}}^{(2)}\|_2}{\sqrt{\|w_N^{(1)}\|_2^2 + \min\limits_{i \in \mathbb{I}_1^{n_2}} |[w_N^{(2)}]_i|^2}}.
\end{aligned}$$

These inequalities, together with the last assumption in Lemma III.1, imply (44). This concludes the proof.

References

[1] P. Eykhoff, *System Identification: Parameter and State Estimation*. Johns Wiley & Sons, 1974.

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

[3] T. Söderström and P. Stoica, *System Identification*. Hertfordshire, United Kingdom: Prentice Hall, 1989.

[4] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*. New York: IEEE Press, 2001.

[5] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, pp. 1–12, 2010.

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

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

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

[9] A. Chiuso and G. Pillonetto, "A Bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48(8), pp. 1553–1565, 2012.

[10] 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.

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

[12] 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 B*, vol. 69 (part 1), pp. 63–78, 2007.

[13] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang, "Subset selection for vector autoregressive processes using Lasso," *Computational Statistics & Data Analysis*, vol. 52, pp. 3645–3657, 2008.

[14] C. R. Rojas and H. Hjalmarsson, "Sparse estimation based on a validation criterion," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC11)*, Orlando, USA, 2011.

[15] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[16] Y. Zhu, "A Box-Jenkins method that is asymptotically globally convergent for open loop data," in *Proceedings of the 18th IFAC World Congress*, Milano, Italy, 2011, pp. 9047–9051.

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

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

[19] 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.

[20] 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.

[21] M. C. Grant and S. P. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control (tribute to M. Vidyasagar)*, V. D. Blondel, S. P. Boyd, and H. Kimura, Eds. Springer-Verlag, 2008, pp. 95–110.

[22] K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems," *IEEE Transactions on Automatic Control*, vol. 10(10), pp. 461–464, October 1965.

[23] P. Stoica and T. Söderström, "The Steiglitz-McBride identification algorithm revisited - convergence analysis and accuracy aspects," *IEEE Transactions on Automatic Control*, vol. 26(3), pp. 712–717, 1981.

[24] P. Regalia, *Adaptive IIR Filtering in Signal Processing and Control*. New York: Marcel Dekker, 1995.

[25] U. Forssell and H. Hjalmarsson, "Maximum likelihood estimation of models with unstable dynamics and non-minimum phase noise zeros," in *Proceedings of the 14th IFAC World Congress*, Beijing, China, 1999.

[26] C. K. Sanathanan and J. Koerner, "Transfer function synthesis as a ratio of two complex polynomials," *IEEE Transactions on Automatic Control*, vol. 8, no. 1, pp. 56–58, January 1963.

[27] 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.

[28] 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.

[29] H. Leeb and B. M. Pötscher, "Sparse estimators and the oracle property, or the return of hodges' estimator," *Journal of Econometrics*, vol. 142, pp. 201–211, 2008.

[30] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, vol. 92(4), pp. 937–950, 2005.

[31] C. R. Rojas, J. C. Agüero, J. S. Welsh, and G. C. Goodwin, "On the equivalence of least costly and traditional experiment design for control," *Automatica*, vol. 44(11), pp. 2706–2715, 2008.

[32] 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.

[33] L. Ljung, *System Identification Toolbox, for use with Matlab*. The Mathworks Inc., 2006.

[34] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[35] K. L. Chung, *A Course in Probability Theory, 3rd Edition*. Academic Press, 2001.

[36] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*. New York: Springer-Verlag, 1985.

[37] L. Ljung and B. Wahlberg, "Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra," *Adv. Appl. Prob.*, vol. 24, pp. 412–440, 1992.

[38] A. N. Kolmogorov and S. V. Fomin, *Introductory Functional Analysis*. Dover, 1975.

[39] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

Cristian R. Rojas (M'13) was born in 1980. He received the M.S. degree in electronics engineering from the Universidad Técnica Federico Santa Maria, Valparaíso, Chile, in 2004, and the Ph.D. degree in electrical engineering at The University of Newcastle, NSW, Australia, in 2008. Since October 2008, he has been with the Royal Institute of Technology, Stockholm, Sweden, where he is currently an Assistant Professor of the Automatic Control Lab, School of Electrical Engineering. His research interests lie in system identification and signal processing.

Roland Tóth was born in 1979 in Miskolc, Hungary. He received the B.Sc. degree in electrical engineering and the M.Sc. degree in information technology in parallel at the University of Pannonia, Veszprém, Hungary, in 2004, and the Ph.D. degree (cum laude) from the Delft Center for Systems and Control (DCSC), Delft University of Technology (TUDelft), Delft, The Netherlands, in 2008. He was a Post-Doctoral Research Fellow at DCSC, TUDelft, in 2009 and at the Berkeley Center for Control and Identification, University of California, Berkeley, in 2010. He held a position at DCSC, TUDelft, in 2011-12. Currently, he is an Assistant Professor at the Control Systems Group, Eindhoven University of Technology (TU/e). He is an Associate Editor of the IEEE Conference Editorial Board. His research interest is in linear parameter-varying (LPV) and nonlinear system identification, modeling and control, machine learning, process modeling and control, and behavioral system theory.

Dr. Tóth received the TUDelft Young Researcher Fellowship Award in 2010.

Håkan Hjalmarsson (M'98, SM'11, F'13) was born in 1962. He received the M.S. degree in Electrical Engineering in 1988, and the Licentiate degree and the Ph.D. degree in Automatic Control in 1990 and 1993, respectively, all from Linkping University, Sweden. He has held visiting research positions at California Institute of Technology, Louvain University and at the University of Newcastle, Australia. He has served as an Associate Editor for Automatica (1996-2001), and IEEE Transactions on Automatic Control (2005-2007) and been Guest Editor for European Journal of Control and Control Engineering Practice. He is Professor at the School of Electrical Engineering, KTH, Stockholm, Sweden. He is an IEEE Fellow and Chair of the IFAC Coordinating Committee CC1 Systems and Signals. In 2001 he received the KTH award for outstanding contribution to undergraduate education. He is co-recipient of the European Research Council advanced grant. His research interests include system identification, signal processing, control and estimation in communication networks and automated tuning of controllers.