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 \( \ell_0 \) “norm” of the parameters. Due to the non-convex nature of the \( \ell_0 \)-norm, this penalization is often implemented as solving an optimization program based on a convex relaxation (e.g., \( \ell_1 \)-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 \( \ell_1 \) 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 \( \ell_2 \) 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.

Index Terms—AIC, BIC, cross-validation, LASSO, model structure selection, sparse estimation, Steiglitz-McBride method, system identification.

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]–[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 \( \ell_1 \) norm of the parameter.
vector (which is known to enforce sparsity). More precisely the LASSO criterion is
\[
\min_{\theta \in \mathbb{R}^n} \quad V(\theta, D_N) \quad \text{subject to} \quad \|\theta\|_1 \leq \varepsilon
\]
where \( V \) is the least-squares cost function based on a data set \( D_N \) 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 problem of minimizing \( V_N(\theta) \) under a constraint on the size of the support of \( \theta \).

Integral to many of the approaches is the use of cross-validation or some information criterion, e.g., the Akaike Information Criterion (AIC) or Generalized Cross-Validation (GCV). For example, such methods can be used to determine the constant \( \varepsilon \) in (1). This means solving (1) and then evaluating the performance of the estimate using, e.g., GCV, for different values of \( \varepsilon \) and then picking the best \( \varepsilon \). While different search strategies for the best \( \varepsilon \) can be derived, a drawback is that it is necessary to solve (1) multiple times [12], [13]. For large problems, this can be prohibitive. 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 \( \varepsilon \) 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, Ch. 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, Ch. 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 \( \ell_1 \) 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 prefiltering 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, \( \|x\|^2_W := x^\top W x \) for \( W = W^\top > 0 \), \( \|x\|^2 := x^\top x \) and \( \|x\|_1 := \sum_{i=1}^n |x_i| \) with \( x = [x_1 \ldots x_n]^\top \). \( \text{Cond}(A) \) is the condition number of a matrix \( A \) in the 2-norm, i.e., \( \text{Cond}(A) := \|A\|\|A^{-1}\| \) where \( \|A\| \) denotes the maximum singular value of \( A \). Notice that \( \text{Cond}(A) = \text{Cond}(A^\top) \geq 1 \). \( \mathbb{P}^n := \{i \in \mathbb{Z} \mid s_1 \leq i \leq s_2 \} \) denotes an index set. The vector containing the signs of a vector \( x \in \mathbb{R}^n \) (in terms of values \( \pm 1 \)) is denoted by \( \text{Sgn}(x) \), while the support of \( x \) is denoted by \( \text{Supp}(x) := \{i \in \mathbb{P}^n \mid |x_i| \neq 0\} \). For a given \( T \subset \mathbb{P}_n \), \( x_T \) denotes the projection of \( x \) to the support \( T \), i.e., \( x_T(i) := |x_i| \) if \( i \in T \) and 0 otherwise.

\( X_N \overset{\text{d}}{\rightarrow} X \) 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 \( A_N/B_N \overset{\text{d}}{\rightarrow} 0 \), and \( A_N \propto_p B_N \) 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 \( N \geq N_0(\varepsilon) \), \( P[m(\varepsilon) < |A_N|/|B_N| < M(\varepsilon)] \geq 1 - \varepsilon \). \( x_N \in \mathbb{N} \) 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 \( y_N \rightarrow y \)) 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

\[
y_t = B_0(q) A_0(q) u_t + C_0(q) D_0(q) c_t
\]
where \( c_t \) is a Gaussian white noise sequence of zero mean and variance \( \sigma^2 > 0 \), \( u_t \) is a quasi-stationary signal [2], and

\[
A_0(q) = 1 + a_{01} q^{-1} + \cdots + a_{0n} q^{-n}
\]
\[
B_0(q) = b_{01} q^{-1} + \cdots + b_{0n} q^{-n}
\]
\[
C_0(q) = 1 + c_{01} q^{-1} + \cdots + c_{0n} q^{-n}
\]
\[
D_0(q) = 1 + d_{01} q^{-1} + \cdots + d_{0n} q^{-n}
\]
with \( q \) the time-shift operator, \( \theta_o = [a_1^o \ldots a_{n_a}^o b_1^o \ldots b_{n_b}^o] \) and \( \eta_o = [c_1^o \ldots c_{n_c}^o d_1^o \ldots d_{n_d}^o] \). Due to physical insights or simply to the generality of the representation, we assume as prior knowledge that only few of the parameters \( \theta_o \) are actually nonzero. 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

\[
y_t = \frac{B(q)}{A(q)} u_t + \frac{C(q)}{D(q)} \epsilon_t
\]

where

\[
A(q) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}
\]

\[
B(q) = b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}
\]

\[
C(q) = 1 + c_1 q^{-1} + \cdots + c_{n_c} q^{-n_c}
\]

\[
D(q) = 1 + d_1 q^{-1} + \cdots + d_{n_d} q^{-n_d}
\]

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 \( D(q) = A(q), \, D_o(q) = A_o(q), \) and \( C(q) = C_o(q) = 1 \). 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

\[
Y_N = \Phi_N \theta^o + E_N
\]

where \( \theta^o \in \mathbb{R}^{n_x} \), \( E_N \sim N(0, \sigma^2 I_N) \) (with \( \sigma^2 > 0 \), \( \Phi_N \in \mathbb{R}^{N \times n_x} \) and \( Y_N \in \mathbb{R}^N \)). Furthermore, \( \Theta_o := \text{Supp} (\theta^o) \) where \( \Theta_o \) contains the indexes (positions) of the non-zero elements of \( \theta^o \), while \( \Theta_a := \mathbb{I}_n^o \setminus \Theta_o \) denotes the positions of the zeros. The model chosen to capture the dynamics of (6) is

\[
Y_N = \Phi_N \theta + E_N
\]

where \( \theta \in \mathbb{R}^{n_x} \) is unknown (which also means that \( \Theta_a \) is a priori unknown). With respect to \( \Phi_N \) and its relation to \( E_N \), we will assume that:

1) \( N^{-1} \Phi_N^T \Phi_N \overset{p}{\rightarrow} \Gamma > 0 \) as \( N \rightarrow \infty \);
2) \( V(\hat{\theta}_N^{LS}, \mathcal{D}_N) \overset{p}{\rightarrow} \sigma^2 \) as \( N \rightarrow \infty \);
3) \( \sqrt{N} (\hat{\theta}_N^{LS} - \theta^o) \in \mathcal{N}(0, \sigma^2 M) \), where \( M \) is a non-singular matrix.

Here,

\[
V(\theta, \mathcal{D}_N) := \frac{1}{N} (Y_N - \Phi_N \theta)^\top (Y_N - \Phi_N \theta)
\]

is the least-squares cost (\( \ell_2 \)-loss of the prediction), and

\[
\hat{\theta}_N^{LS} := (\Phi_N^T \Phi_N)^{-1} \Phi_N^T Y_N
\]

is the least-squares estimate.

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

\[
A(q) y_t = B(q) u_t + \epsilon_t
\]

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 \( Y_N := [y_{n_a+1} \ldots y_N]^\top \), \( E_N := [\epsilon_{n_a+1} \ldots \epsilon_N]^\top \), \( \theta := [a_1 \ldots a_{n_a} b_1 \ldots b_{n_b}] \) and

\[
\Phi_N = \begin{bmatrix}
-y_{n_a} & \cdots & -y_1 & u_{n_a} & \cdots & u_{n_a-n_b+1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
-y_{N-1} & \cdots & -y_{N-n_a} & u_{N-1} & \cdots & u_{N-n_b}
\end{bmatrix}
\]

(For the sake of simplicity, we assume that \( n_a > n_b \).)

**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 \( \theta^o \) for \( N \geq n_g \). 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 \( \Phi_N \) is deterministic and satisfies \( -\Phi_N^T \Phi_N \rightarrow 0 \) as \( N \rightarrow \infty \).

2) **Method:** The method we propose for estimating a sparse \( \theta \) is based on the following steps:

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

ii) Obtain a sparse estimate \( \theta_N \) by solving

\[
\min_{\theta \in \mathbb{R}^{n_x}} \| \theta \|_1 \quad \text{s.t.} \quad V(\theta, \mathcal{D}_N) \leq V(\hat{\theta}_N^{LS}, \mathcal{D}_N) (1 + \varepsilon_N)
\]

where \( \varepsilon_N > 0 \) and \( n_b := n_a + n_b \). The choice of \( \varepsilon_N \) will be discussed later.

iii) Finally, re-estimate the non-zero elements of \( \theta_N \) using ordinary least-squares. More precisely, let \( \mathcal{T} \) correspond to the indexes of the non-zero elements in \( \theta_N \). Define \( \Phi_{N,T} \) to be the matrix formed from the columns of \( \Phi_{N} \) listed in \( \mathcal{T} \) and then compute the least-squares estimate of a \( \theta \) of reduced dimension based on the model (7) with \( \Phi_{N,T} \). 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 $\hat{\theta}_N^{\text{RE}}$.

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

$$
\begin{align}
\min_{\theta \in \mathbb{R}^{r \times n}} & \|w_N \otimes \theta\|_1 \\
\text{s.t.} & \quad V(\theta, D_N) \leq V\left(\hat{\theta}_{N}^{\text{LS}}, D_N\right) (1 + \varepsilon_N)
\end{align}
$$

(13a)

(13b)

where $w_N \in \mathbb{R}^{r \times n}$ is given by $w_{N,i} := 1/|\hat{\theta}_{N,i}^{\text{LS}}|$ with $i \in I^r \times n$ 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 $\hat{\theta}_N^A$; the method with all three steps, in this case, is denoted as A-SPARSEVA-RE and the corresponding estimate is $\hat{\theta}_N^{A-\text{RE}}$. 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 $\varepsilon_N$ 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 $D_N = \{u_t, y_t\}_{t=1}^N$ of (2) and the model structure (10) characterized by the parameters $\theta = [a_1 \ldots a_m] \in \Theta \subseteq \mathbb{R}^{r \times n}$. Assume that $D_N$ is informative w.r.t. (10), see [2].

1: Compute $\hat{\theta}_N^{\text{LS}}$ via (9).
2: Set $\varepsilon_N = 2n_g / N$ or $\varepsilon_N = (n_g \log N) / N$ and compute $V\left(\hat{\theta}_N^{\text{LS}}, D_N\right)$.
3: Obtain the sparse estimate $\hat{\theta}_N^A$ by solving (13). (In the non-adaptive case, solve (12) to obtain $\hat{\theta}_N^A$).
4: Based on a threshold $0 \leq \varepsilon_* < \varepsilon_N$, select a minimal $T \subseteq \text{Supp}(\hat{\theta}_N^{A,T})$ such that $\|\hat{\theta}_N^{A,T} - \hat{\theta}_N^A\|_1 \leq \varepsilon_* \|\hat{\theta}_N^A\|_1$.  
5: Estimate $\hat{\theta}_N^{A-\text{RE}}$ via (9) with $\Phi_{N,T}$.  
6: return estimated model (10).

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

$$
E_{\text{val}}\left\{E_{\text{est}}\left\{V\left(\hat{\theta}_N^{\text{LS}}, D_N^{\text{val}}\right)\right\}\right\} = \left(1 + \frac{2n_g}{N}\right)E_{\text{est}}\left\{V\left(\hat{\theta}_N^{\text{LS}}, D_N\right)\right\}
$$

(14)

where $E_{\text{est}}\{\cdot\}$ ($E_{\text{val}}\{\cdot\}$) 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

$$
\left(1 + \frac{2n_g}{N}\right) V\left(\hat{\theta}_N^{\text{LS}}, D_N\right)
$$

(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 $\varepsilon_N = 2n_g / N$, (12) can be seen as a way to estimate a sparse (due to the $\ell_1$-norm) model such that its performance is similar to $V\left(\hat{\theta}_N^{\text{LS}}, D_N^{\text{val}}\right)$. Thus, unlike for the LASSO, there is a natural choice of the “regularization” parameter $\varepsilon_N$ for SPARSEVA, which corresponds to a particular level set of the $\ell_2$-loss function (8), in which the loss of the sparse solution is expected to lie. In the LASSO case, the $\ell_2$ cost of the prediction error is minimized for a given sparsity level, i.e., $\|\hat{\theta}\|_1 < \varepsilon$, see (1). As the $\ell_1$ norm of the optimal estimate for $\theta$ is generally unknown, it is much harder in practice to develop a selection scheme for $\varepsilon$ in the LASSO method. This is the motivation for introducing (12).

It should be noted that the convex optimization program

$$
\begin{align}
\min_{\theta} & \|\theta\|_1 \\
\text{s.t.} & \quad V(\theta, D_N) \leq \varepsilon
\end{align}
$$

(16a)

(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 $\varepsilon_N$ according to (12), in particular with $\varepsilon_N$ chosen by the AIC rule $\varepsilon_N = 2n_g / N$, and in the adaptive version (13) inspired by [18].

The use of a threshold $\varepsilon_*$ 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 $\hat{\theta}_N^{A,k}$ which is sparse only up to some numerical precision (e.g., $10^{-10}$). The choice of $\varepsilon_*$ should be made according to the precision of the 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 $\varepsilon_*$ is negligible1 (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,

$$
y_t = \frac{B(q)}{A(q)} u_t + e_t
$$

(17)

which corresponds to (4) with $C(q) = D(q) = 1$. It is well known, see [2], that the least-squares estimator $\hat{\theta}_N^{LS} := \left(\Phi_N^{\top} \Phi_N\right)^{-1} \Phi_N^{\top} Y_N$ (where $\Phi_N$ 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 $1/A(k)^{(q)}$ resulting in the filtered data set $D_N^{(k)}$, where $A(k)^{(q)}$ is an estimate of the $A(q)$ polynomial (at step $k$).

---

1Many 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.
Then, least-squares estimation is applied on $D_N^{(k)}$, 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

$$A_{HO}(q)y_t = B_{HO}(q)u_t + \epsilon_t$$

is fitted to $D_N$, and used then to pre-filter the data, i.e., to generate the signals

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

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 $1/\hat{A}_{HO}(q)$ should be a reasonable estimate of the noise model $C_n(q)/D_n(q)$, 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 $\hat{A}(q)$ and $\hat{A}_{HO}$: 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 $\hat{A}(q)$ as $\hat{A}_+(q)\hat{A}_-(q)$, where $\hat{A}_+$ is stable and $\hat{A}_-$ is anti-stable (the constant factor can be arbitrarily assigned to any of these factors), pre-filter the data forward in time using $1/\hat{A}_+(q)$, and then use the filter $1/\hat{A}_-(q)$ 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, SPRASEVA and $\ell_1$-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 $\ell_1$-penalized estimators in order to deliver sparse models.

\begin{algorithm}[H]
\caption{OE-SPARSEVA with Steiglitz-McBride}
\begin{algorithmic}[1]
\REQUIRE a data record $D_N = \{u_i, y_i\}_{i=1}^N$ of (2) and the model structure (17) characterized by the parameters $\theta = [a_1 \ldots b_{m_b}]^T \in \Theta \subseteq \mathbb{R}^{n_a+n_b}$. Assume that $D_N$ is informative w.r.t. (17) and (17) is globally identifiable on $\Theta$ [2].
\STATE Let $m \gg n_a$ and fit using least-squares the high order ARX model described by (18) to the measurements $D_N$, resulting in $\hat{A}_{HO}(q)$ and $\hat{B}_{HO}(q)$.
\STATE Filter the data $D_N$ as

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

\STATE Set $k = 0$, and let $\hat{A}^{(0)}(q) = 1, \quad \hat{B}^{(0)}(q) = 0$ and consequently $\hat{\theta}_{(0)} = 0$.
\FOR{$k \leftarrow k + 1$}
\STATE Filter the data $D_N$ as

$$y_t^{F(k)} := \frac{1}{A^{(k-1)}(q)}y_t^F, \quad u_t^{F(k)} := \frac{1}{A^{(k-1)}(q)}u_t^F.$$  

\STATE Fit, using least-squares, a model of the form

$$A^{(k)}(q)y_t^{F(k)} = B^{(k)}(q)u_t^{F(k)} + \epsilon_t^{(k)}$$  

resulting in the estimates $\hat{A}^{(k)}$, $\hat{B}^{(k)}$ and the associated parameter vector $\hat{\theta}_{(k)}^{(0)}$.
\UNTIL $\hat{\theta}_{(k)}$ has converged or the maximum number of iterations is reached.
\STATE Apply A-SPARSEVA (with least-squares re-estimation) to the model

$$A(q)y_t^{F(k+1)} = B(q)u_t^{F(k+1)} + \epsilon_t^{(k+1)}.$$  

\RETURN estimated model (17).
\end{algorithmic}
\end{algorithm}

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 $\ell_1$-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 $\varepsilon_N$ can also be applied to recover the exact sparsity structure of $\theta$. 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 $V(\hat{\theta}_{N}^{(0)}, D_N)$ 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 SPRASEVA scheme is an automatic choice of $\varepsilon_N$ guaranteeing a reliable performance.
w.r.t. (6), we have the following results: 

\[ \theta \]

assumptions of Section IV-A-1, and 

Fig. 1. Geometry of (12). In (a), a sparse solution (ROJAS et al.) is obtained but not in (b).

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 \( \varepsilon^* = 0 \).

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-A-1, and \( \theta^0 \neq 0 \), the SPARSEVA and A-SPARSEVA estimators are consistent in probability (i.e., \( \hat{\theta}^N \to \theta^0 \)) if and only if \( \varepsilon \to 0 \). In particular, \( \| \hat{\theta}^N(\varepsilon) - \theta^0 \|_2 = O_p(N^{-1/2} + \sqrt{\varepsilon N}) \) uniformly in \( \theta^0 \).

**Proof:** See Appendix B.

**Corollary 5.1 (Exact Order of Consistency):** Subject to the assumptions of Theorem 5.1, if \( \varepsilon \to 0 \), but \( N \varepsilon \to \infty \), then \( \| \hat{\theta}^N(\varepsilon) - \theta^0 \|_2 \geq \sqrt{\varepsilon N} \) (c.f. Section II for the definition of \( \approx_p \)).

**Proof:** See Appendix C.

2) Sparsity: Since \( V(\hat{\theta}, D_N) \) is quadratic, the constraint (12b) corresponds to an ellipsoid \( \Omega \) in \( \Theta \subseteq \mathbb{R}^n \). The solution to (12a) will be on the boundary of the smallest \( \ell_1 \)-ball that intersects this ellipsoid, see Fig. 1(a). When the ellipsoid has a shape as in Fig. 1(a), then, as can be seen, the solution will be sparse. However, with a more tilted ellipsoid as in Fig. 1(b), the solution will not be sparse. The shape of the ellipsoid is determined by the regressor matrix \( \Phi_N \).

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 \( \varepsilon_N \) for the adaptive SPARSEVA to generate sparse estimates (recovery of the true support of \( \theta^0 \)).

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

**Proof:** See Appendix D.

**Remark 5.1:** It can be shown that when the regressors are orthonormal, i.e., \( N^{-1} \Phi_N^T \Phi_N = I \), 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 \( \varepsilon_N \) 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 \( \varepsilon_N \) gives rise to a non efficient estimator (since the order of convergence of \( \hat{\theta}^N \) to \( \theta^0 \) 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,

\[ \sqrt{N} \left( \hat{\theta}^N \to \theta^0 \right) \in \mathcal{N}(0, M^{-1}) \]

where \( M \) is the information matrix when it is known which elements of \( \theta^0 \) 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 \( \hat{\theta}^N \) as an estimator of the location of the non-zero components of \( \theta^0 \) by any other \( \varepsilon \)-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 \( \theta^0 \) such that the risk \( R(\hat{\theta}_N, \theta^0) := E(\| \hat{\theta}_N - \theta^0 \|_2) \) is small. Since the risk \( R \) depends on the unknown true value \( \theta^0 \), it is relevant to study the worst-case performance, \( \sup_{\theta^0 \in \mathbb{R}^n} R(\hat{\theta}_N, \theta^0) \). 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 \( \theta^0 \), if \( \sup_{\theta^0 \in \mathbb{R}^{\varepsilon N}} R(\hat{\theta}_N, \theta^0) \) converges to zero at the same rate as \( \inf_{\hat{\theta}_N} \sup_{\theta^0 \in \mathbb{R}^n} R(\hat{\theta}_N, \theta^0) \), where \( \delta \) ranges over all estimators of \( \theta^0 \) 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-A-1, A-SPARSEVA (13) and \( \hat{\theta}^N_{A-RE} \) 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 trade-off 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

\[ \text{Fig. 1. Geometry of (12). In (a), a sparse solution (\theta_2 = 0) is obtained but not in (b).} \]
functions, such as the $\ell_1$ 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 loses its minimax rate optimality.

Remark 5.5: Notice that the scaling of the parameters in $\theta_0$ 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 $\ell_1$ norm by the inverse of the estimates in $\hat{\theta}_N^{LS}$, which compensates for the relative size of the components of $\theta^o$.

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 $\varepsilon_N$, 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 prefilter 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 $A_n(q)$ and $B_n(q)$ for general Box-Jenkins model structures3 (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 $\theta^o \neq 0$, 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]:

$$\min_{\theta \in \mathbb{R}^{n_s}} \quad V(\theta, D_N)$$

s.t. \quad $\|w_N \circ \theta\|_1 \leq \varepsilon_L$.

This estimator can be written in the Lagrangian form

$$\Lambda_L(\theta, \lambda_L) = V(\theta, D_N) + \lambda_L(\|w_N \circ \theta\|_1 - \varepsilon_L) \quad (23)$$

with $\lambda_L \geq 0$. The optimum of (22) is obtained at the optimum of

$$\max_{\lambda_L \geq 0} \Lambda_L(\theta, \lambda_L). \quad (24)$$

Similarly, (13) has the Lagrangian:

$$\Lambda_S(\theta, \lambda_S) = \|w_N \circ \theta\|_1 + \lambda_S(V(\theta, D_N) - V(\hat{\theta}_N^{LS}, D_N)(1 + \varepsilon_N)) \quad (25)$$

Notice that both $V(\cdot, D_N)$ and $\|\cdot\|_1$ are convex functions, $V(\cdot, D_N)$ is strongly convex and all constraints satisfy the constraint qualification, hence solutions of (24) and (25), i.e., $(\theta_L(\varepsilon_L), \lambda_L^*(\varepsilon_L))$ and $(\theta_S(\varepsilon_S), \lambda_S^*(\varepsilon_S))$ are unique with no duality gap.

For a given $\varepsilon_L$, let $\varepsilon_N$ be such that $V(\hat{\theta}_N^{LS}, D_N)(1 + \varepsilon_N)$ is equal to the minimum value of (22). For this choice of $\varepsilon_N$, the feasibility sets $U_L := \{\theta \in \mathbb{R}^{n_s} : \|w_N \circ \theta\|_1 \leq \varepsilon_L\}$ and $U_S := \{\theta \in \mathbb{R}^{n_s} : V(\theta, D_N) \leq V(\hat{\theta}_N^{LS}, D_N)(1 + \varepsilon_N)\}$ are convex and intersect at exactly one point,$^4$ $\theta_L(\varepsilon_L) = \theta_S(\varepsilon_N) = \theta_L(\varepsilon_L)$.

Notice, however, that this relation (which is induced by the KK.T conditions of (23) and (25)) is dependent on the optimal solution $\theta^o$, i.e., it is data-dependent (the center of $U_S$ depends on $\hat{\theta}_N^{LS}$ and its shape depends on $\Phi_N$).

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

$$\min_{\bar{w} \in \mathbb{R}^{n_g}} \quad V(\bar{w} \circ \hat{\theta}_N^{LS}, D_N) + \lambda_N\|\bar{w}\|_1$$

s.t. $\bar{w} \geq 0$.

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

$$\min_{\theta \in \mathbb{R}^{n_s}} \quad V(\theta, D_N) + \lambda_N\|w_N \circ \theta\|_1$$

s.t. $\theta \circ \hat{\theta}_N^{LS} \geq 0$.

as $[w_N]_i := 1/\|\hat{\theta}_N^{LS}\|_i$. 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 $\lambda_N$ and the optimal $\lambda_L^*$).

$^4$If $\theta_S \neq \theta_L$ both belong to $U_S \cap U_L$, then both achieve the minimum of (22). However, $\theta_S + \theta_L/2$ also belongs to $U_S \cap U_L$ (since it is a convex set), and due to the strong convexity of $V$, $\|w_N \circ (\theta_S + \theta_L/2)\|_1 < V(\theta_S, D_N) < V(\theta_L, D_N)$, contradicting the optimality of $\theta_S$ and $\theta_L$. This means that the optimum of (22) is unique and that $U_S \cap U_L$ is a singleton.
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 \( \theta^0 \))- 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,

\[
A_0(q) = 1, \quad B_o(q) = 3q^{-1} + 1.5q^{-2} + 2q^{-5}
\]

\[
C_o(q) = 1, \quad D_o(q) = A_o(q) = 1.
\]

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 \( n_a = 0 \) and \( n_b = 8 \), which results in the true parameter vector

\[
\theta^0 = [3 \ 1.5 \ 0 \ 0 \ 2 \ 0 \ 0 \ 0]^	op.
\]

Notice that \( \theta^0 \) 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 \( \Phi_N \) 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 \( \Phi_N \), 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 \( w_t \sim N(0, 1) \).

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 \( 11 \times 100 \) estimation and validation data records. The average Signal to Noise Ratio\(^5\) (SNR) has been \(-3.97 \text{ dB} \).

Using these data sets, SPARSEVA is compared to the following methods: LS-ORACLE: least-squares estimate of \( \theta \) using \( \Phi_{\mathcal{T}_N} \) (prior knowledge of the non-zero parameters). Note that

\[
\text{SNR} = 10 \cdot \log_{10}(\frac{\|y_t - v_t\|^2}{\|v_t\|^2}) \quad \text{where} \quad v_t = (C_o(q)/D_o(q))w_t.
\]

Fig. 2 shows the Mean-Squared Error (MSE) of the parameter estimate as a function of the sample size \( N \).

\[\text{MSE of parameter estimate} \]

\[\text{LS-ORACLE} \quad \text{LASSO-GCV} \quad \text{SPARSEVA-AIC/RE} \quad \text{A-SPARSEVA-AIC/RE} \quad \text{A-SPARSEVA-BIC/RE} \]

\[\text{Fig. 2. MSE as a function of the sample size N.} \]

\[\text{this is the ideal estimator and by the Cramér-Rao lower bound w.r.t. the true support of } \theta^0 \text{, no other estimator can perform better. However, it cannot be applied in practice as the optimal model structure is unknown. LASSO-GCV is the LASSO,} \]

\[
\min_{\theta \in \mathbb{R}^p} \quad V(\theta, D_N) + \lambda ||\theta||_1
\]

\[\text{where the regularization parameter } \lambda \text{ is chosen according to generalized cross-validation [32], i.e., the } \lambda \text{ that minimizes} \]

\[
V(\theta^0, D_N) \quad \frac{1}{(1-p(\lambda)/N)^2}
\]

\[\text{is chosen. Here } p(\lambda) \text{ is the number of effective parameters defined as} \]

\[
p(\lambda) = \text{Tr} \left\{ \Phi_N \left( \Phi_N^\top \Phi_N + \lambda W^\top \right)^{-1} \Phi_N^\top \right\}
\]

\[W = \text{Diag} \left( \left| \theta^0 \right| \right) \]

\[\text{with } | \cdot | \text{ taken element-wise. Four variants of SPARSEVA are included: SPARSEVA-AIC/BIC where the constraint } \varepsilon_N \text{ is chosen as AIC } (\varepsilon_N = 2n_R/N) \text{ and BIC } (\varepsilon_N = (n_R \log N)/N). \text{ A-SPARSEVA-AIC/BIC are the two corresponding adaptive versions. Notice that the BIC choice for } \varepsilon_N \text{ satisfies the condition for sparseness (see Theorem 5.2).} \]

\[\text{Under these conditions, 100 estimation and 100 validation data sets have been generated for each “data length” } N \in \{10 + 10k\}_{k=1}^{10} \text{ resulting in } 11 \times 100 \text{ estimation and validation data records. The average Signal to Noise Ratio}\(^5\) (SNR) has been } -3.97 \text{ dB}. \]

\[\text{Using these data sets, SPARSEVA is compared to the following methods: LS-ORACLE: least-squares estimate of } \theta \text{ using } \Phi_{\mathcal{T}_N} \text{ (prior knowledge of the non-zero parameters). Note that} \]

\[\text{\footnote{The collapse of the two MSE curves is due to the inherent randomness of the Monte Carlo simulations, not to the equality of the actual MSE of the estimators for } N \geq 70.} \]
parameters, and we see that this estimator has the best ability to determine where the zero elements are located. However, from Fig. 2 it can be seen that for small sample sizes the performance of this estimator is worse than for almost all other estimators. From Fig. 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 $\hat{\theta}^p\|_2$ are zero. The results are compared in terms of

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

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

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

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

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

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 $\ell_1$ parameter estimation error: $\| \hat{\theta} - \theta_o \|_1$.

- The percentage of correctly estimated zero elements.

The average results of the 100 Monte-Carlo runs in each cases is given in Table I. From these results it follows that in the low noise cases (SNR = 30 dB, $N = 2000$ case) the SMB-ORACLE estimator in terms of the Steiglitz-McBride method has been also applied with the priori knowledge of which elements of $\theta^p\|_2$ are zero. The parameters have no role at all in the considered model structure.

When the noise increases to a moderate level (SNR = 7.5 dB), for small data lengths we can observe that OE-SPARSEVA has the best ability to determine where the zero elements are located.
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 approach has a much slower convergence rate. The same behavior can be observed in the SNR = 0 dB 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. Conclusion

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 \( \varepsilon_N = (1 + 2\eta/N)V(\theta^{LS}_N, D_N) \) seems to give a good balance between sparsity and model fit. Thus, this method \( \varepsilon \) performs most favorably. On these examples, the “AIC” simulations have shown that the adaptive version of this method

\[
\begin{array}{c|c|c|c|c|c|c|c}
\text{Method} & a_2 & a_8 & b_6 & b_7 \\
\hline
\theta^0 & -0.1972 & -0.2741 & 1 & -8.3365 \\
SMB-ORACLE mean & -0.1976 & -0.2714 & 1.0028 & -8.3535 \\
standard deviation & 0.0140 & 0.0169 & 0.1818 & 0.1656 \\
OE mean & -0.1975 & -0.2737 & 0.9937 & -8.3490 \\
standard deviation & 0.0155 & 0.0157 & 0.1862 & 0.1652 \\
OE-SPARSEVA mean & -0.1972 & -0.2713 & 0.9985 & -8.3507 \\
standard deviation & 0.0140 & 0.0166 & 0.1794 & 0.1676 \\
\end{array}
\]

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

\[
\begin{align}
\min_{\theta \in \mathbb{R}^n} & \quad \|\omega & \odot \theta\|_1 \quad (34a) \\
\text{s.t.} & \quad \sigma^2 \varepsilon_N \eta \geq \|\theta - (\theta^0 + \xi)\|_N^2 \quad (34b)
\end{align}
\]

where \( \Gamma_N := \Phi_N^\top \Phi_N, \omega = [1 \ldots 1] \in \mathbb{R}^n \) for (12) and \( \omega = w_N \) for (13).

**Proof:** See Appendix G.

To simplify the notation, define

\[
\Omega := \left\{ \theta \in \mathbb{R}^n : \|\theta - (\theta^0 + \xi)\|_{\Gamma_N} \leq \sqrt{\sigma^2 \varepsilon_N \eta} \right\}
\]

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 \( \Omega \), and no nonzero interior point of \( \Omega \) can be an optimum point of (34).

**Proof:** See Appendix G.

Since \( \eta/N \overset{p}{\to} 1 \) and since

\[
\frac{1}{N}\|z\|^2_{\Gamma_N} = \frac{1}{N}z^\top \Gamma_N z \geq \left\{ \lambda_{\min}(\Gamma) + o_p(1) \right\} \|z\|^2_2
\]

for every \( z \in \mathbb{R}^n \), (34) gives that

\[
\|\hat{\theta}_N - \theta^0\|^2_2 \leq \|\hat{\theta}_N - \hat{\theta}_{LS}^N\|^2_2 + \|\hat{\theta}_{LS}^N - \theta^0\|^2_2
\]

\[
\leq \frac{\|\hat{\theta}_N - \hat{\theta}_{LS}^N\|_{\Gamma_N}}{\sqrt{N(\lambda_{\min}(\Gamma) - o_p(1))}} + \|\hat{\theta}_{LS}^N - \theta^0\|^2_2
\]

\[
\leq \frac{\sigma^2 \varepsilon_N \eta}{N} \left\{ \lambda_{-1/2}^{-1}(\Gamma) + o_p(1) \right\} + \|\xi\|^2_2
\]

\[
= \sigma \lambda_{-1/2}^{-1}(\Gamma) \sqrt{\varepsilon_N} + o_p(1) + O_p(N^{-1/2})
\]

which implies that \( \hat{\theta}_N \overset{p}{\to} \theta^0 \) if \( \varepsilon_N \to 0 \). Conversely, assume that \( \liminf_{N \to \infty} \varepsilon_N = \delta > 0 \), i.e., that there is a subsequence \( \{N_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) 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 \( 1 - \delta > 1/2 \), \( \|\hat{\theta}_{LS}^N - \theta^0\|^2_2 < \sigma \sqrt{(\delta/12)} |\Gamma|^{-1/2} \) of \( \theta^0 \).

Then, since for all \( k \in \mathbb{N} \) and \( x \in \mathbb{R}^x \), under \( S_{\varepsilon} \), it holds that

\[
\frac{1}{N_k} \|x\|^2_{\Gamma_{N_k}} \leq \frac{1}{N_k} x^\top \Gamma_{N_k} x \leq \|N_k^{-1} \Gamma_{N_k}\| \|x_x\|^2 \leq \frac{3}{2} \|\Gamma\| \|x_x\|^2
\]

and \( \|\theta - (\theta^0 + \xi)\|^2_2 \leq \|\theta - \theta^0\|^2_2 + \|\xi\|^2_2 \), it follows that under \( S_{\varepsilon} \):

\[
\Omega \ni \frac{\sqrt{2\sigma^2 \varepsilon_N \eta}}{N_k} \frac{\sqrt{\varepsilon_N \eta}}{N_k} - \|\theta - \theta^0\|^2_2
\]

\[
\Omega \ni \left\{ \theta \in \mathbb{R}^n : \|\theta - (\theta^0 + \xi)\|_{\Gamma_N} \leq \sqrt{\sigma^2 \varepsilon_N \eta} N_k / N_k \right\}
\]

7The condition \( \lim_{N \to \infty} N^{-1/2} \Gamma_N = \Gamma \) denotes element-wise convergence of the eigenvalues of a matrix being continuous functions of its elements [34, Appendix D], such condition implies the convergence of the eigenvalues of \( N^{-1/2} \Gamma_N \) to the eigenvalues of \( \Gamma \) (appropriately sorted).
This implies that, for all \( k \in \mathbb{N} \),
\[
P\{U \subset \Omega, S_k \} \geq P\left\{ \left\| \xi \right\|_2 < \sqrt{\frac{\sigma^2 \delta}{3 \| \Gamma \|}} \sqrt{\frac{\eta}{N_k}} + \sqrt{\frac{\sigma^2 \delta}{12 \| \Gamma \|}}, S_k \right\} = P\left\{ \left\| \xi \right\|_2 < \sqrt{\frac{\sigma^2 \delta}{12 \| \Gamma \|} + o_p(1)}, S_k \right\} \rightarrow P\{ S_k \} = 1 - \delta
\]
because \( \sqrt{\frac{\eta}{N_k}} \in \mathcal{N}(0, \sigma^2 M) \), i.e., \( \xi = o_p(1) \). Since \( \delta \) was arbitrary, this shows that \( \limsup_{N \rightarrow \infty} P\{U \subset \Omega \} = 1 \), that is, \( \theta = \theta^0 \) is an interior point of the constraint set (34) with non-negligible probability for \( N = N_k \) as \( k \rightarrow \infty \). On the other hand, by Lemma 2.1 the optimum of (34) is achieved on the boundary of \( \Omega \) (or at \( \theta = \theta^0 \)), which means that
\[
\liminf_{N \rightarrow \infty} P\left\{ \left\| \hat{\xi}_N^{(A)} - \theta \right\|_2 > \min \left\{ \sigma \sqrt{\frac{\delta}{12 \| \Gamma \|} + \| \theta^0 \|_2} \right\} > 0
\]
i.e., (A-)SPARSEVA is not consistent in probability if \( \varepsilon_N \rightarrow 0 \).

**Proof of Corollary 5.1**

The corollary follows from the fact (c.f. Lemma 2.2) that the optimum of (34) is achieved on the boundary of the constraint set \( \Omega \) (or at \( \theta = \theta^0 \)), whose size has order \( \varepsilon_N \), larger than \( N^{-1/2} \).

**APPENDIX C**

**PROOF OF THEOREM 5.2**

Without loss of generality, let us assume that \( T_0 = \{1, \ldots, n_1 \} \), with \( n_2 := n_2 - n_1 > 0 \). Note that such a condition can be satisfied by reordering the columns of \( \Phi \) in (6).

Then, the following result holds:

**Lemma III.1 (Conditions for Sparsity):** Under the stated assumptions for (34), and \( \theta^0 \neq 0 \), the optimal solution of (34) (with \( \omega = w_N \)), \( \hat{\theta}_N = \left[ \hat{\theta}_N^{(A, 1)} \right]^\top \left[ \hat{\theta}_N^{(A, 2)} \right]^\top \), with \( \hat{\theta}_N^{(A, i)} \in \mathbb{R}^{n_i} \) (\( i = 1, 2 \)), satisfies \( \text{Supp}(\hat{\theta}_N^{(A, 2)}) = \pi_{n_1} \) and \( \text{Supp}(\hat{\theta}_N^{(A, 1)}) = \emptyset \), i.e., recovery of the true support \( T_0 \) holds, if the following conditions hold:

\[ \Gamma + o_p(1) > 0, \]
\[ \sigma \sqrt{\frac{\lambda_{\min}(\Gamma)}{N}} \min_{i \in \pi_{n_1}} \| \theta^0 \|_i < o_p(1), \]
\[ \sigma \sqrt{\frac{\lambda_{\max}(\Gamma)}{N}} \left\{ 1 + \text{Cond}(\Gamma) \right\} \lambda_{\max}(\Gamma) \left( 1 + \text{Cond}(\Gamma) \right) \lambda_{\max}(\Gamma)^{-1}, \]
\[ \sqrt{\frac{\| \phi \|_2}{\| \phi \|_2}} \leq \lambda_{\min}(\Gamma) \frac{\| \phi \|_2}{\| \phi \|_2} + o_p(1). \]

By Theorem 5.1, \( \| \hat{\theta}_N^{(A, 1)} - \theta^0 \|_2 \leq O_p(N^{-1/2}) \), which implies that \( \| \hat{\theta}_N^{(A, 1)} - \theta^0 + \eta \|_2 \|_2 \leq O_p(N^{-1}). \) Letting \( \hat{\theta}_N^{(A, 1)} \) fixed, it follows from (36) that \( \hat{\theta}_N^{(A, 2)} = 0 \) if and only if \( \hat{\theta}_N = \left[ \theta^0 \right]^\top \) satisfies the constraint in (36), i.e., if
\[
\sigma^2 \varepsilon_N N + o_p(N) \geq \left( \hat{\theta}_N^{(A, 1)} - \theta^0 + \eta \right) \Gamma_N \left( \hat{\theta}_N^{(A, 1)} - \theta^0 + \eta \right) = \lambda_{\min}(\Gamma) N \| \eta \|_2^2 + O_p(1). \]

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

**APPENDIX D**

**PROOF OF THEOREM 5.3**

Denote by \( \hat{\theta}_N^{(A, 1)} \) the least-squares estimate, which is obtained with the exact knowledge of the true support \( T_0 \), i.e., with (9) using \( \Phi_{N,T_0} \). Furthermore, let \( M \) be the asymptotic information matrix of \( \theta \) assuming knowledge of \( T_0 \) and \( S_N \).
be the event that $\text{Supp} (\hat{\theta}_N) = T_n$. The complement of $S_N$ is denoted by $\bar{S}_N$. This gives that, for every $x \in \mathbb{R}^n$, \[
P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x \right\} \]
\[
= P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x | S_N \right\} P \{ S_N \}
+ P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x, \bar{S}_N \right\}
\]
\[
= P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x \right\} P \{ S_N \}
+ P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x \right\} P \{ \bar{S}_N \}
\]
where $\leq$ is taken component-wisely, and $M^{1/2}$ denotes the positive definite square root of $M$. Therefore, \[
\left| P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x \right\} - P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} \right|
\]
\[
= P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} (P \{ S_N \} - 1)
+ P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} P \{ \bar{S}_N \}
\]
\[
\leq 2 P \{ \bar{S}_N \}.
\] (38)
Hence, if $F(x)$ denotes the cumulative standard normal distribution function, then (38) implies that \[
\lim_{N \to \infty} \left| P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{RE}} - \theta^o \right) \right| \leq x \right\} - F(x) \right|
\leq \lim_{N \to \infty} \left| P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} - F(x) \right|
\]
\[
+ \lim_{N \to \infty} \left| P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} - F(x) \right|
\]
\[
\leq 2 \lim_{N \to \infty} P \{ \bar{S}_N \}
\]
\[
\lim_{N \to \infty} \left| P \left\{ \left| [N] \frac{1}{2} \left( \hat{\theta}_N^{\text{oracle}} - \theta^o \right) \right| \leq x \right\} - F(x) \right| = 0
\] (39)
since $\hat{\theta}_N^{\text{oracle}}$ has the sparseness property (i.e., $\lim_{N \to \infty} P \{ \bar{S}_N \} = 0$), and $\hat{\theta}_N^{\text{oracle}}$ is asymptotically efficient and normal. Equation (39) shows that $\hat{\theta}_N^{\text{RE}}$ has the Oracle property.

**APPENDIX E**

**PROOF OF THEOREM 5.4**

First notice that $\inf_{\delta \in \mathbb{R}^n} \sup_{\theta^o \in \mathbb{R}^n} R(\delta, \theta^o) \asymp_p N^{-1}$. To see this, note that the estimator $\hat{\theta}_N^{\text{RE}}$ is minimax optimal because it coincides with the maximum likelihood estimator of $\theta^o$ (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, $\sup_{\theta^o \in \mathbb{R}^n} R(\hat{\theta}_N, \theta^o) = O_p(N^{-1})$, 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 $\mathbb{R}^n$ consisting of all points $\theta \in \mathbb{R}^n$ such that $\theta_{T_A} = 0$ with $T_A = \text{Supp}(\hat{\theta}_N^o)$; 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, $\mathbb{L} \cap \Omega \neq \emptyset$ with high probability, because $\hat{\theta}_N \in \mathbb{L} \cap \Omega$ with high probability. This implies that, for $N$ sufficiently large, $\hat{\theta}_N^{\text{RE}} \in \Omega$ with high probability, since otherwise $V(\hat{\theta}_N^{\text{RE}}, D_N) > V(\hat{\theta}_N^o, D_N)$, contradicting the optimality of $V(\hat{\theta}_N^{\text{RE}}, D_N)$. This observation implies, by following a similar argument as in the proof of Theorem 5.1, that $\sup_{\theta^o \in \mathbb{R}^n} R(\hat{\theta}_N^{\text{RE}}, \theta^o) = O_p(N^{-1})$.

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 $\varepsilon_N$ to a subsequence $\{N_k\}_{k \in \mathbb{N}}$ such that $N_k \to \infty$ 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 $N_k \sup_{\theta^o \in \mathbb{R}^n} R(\hat{\theta}_N^{\text{RE}}, \theta^o) \to \infty$. 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-A-1 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 $A^{(k+1)}(t)$ which is an asymptotically efficient estimate of $A(t)$, i.e., $A^{(k+1)}(t) = A(t) + O_p(N^{-1/2})$. In addition, the data satisfies asymptotically an ARX structure of the form

$A_q(t) v_q = B_q(t) u_q + e_q$

where $v_q = H_q(q) e_q$, with $\{e_q\}$ a Gaussian white noise sequence of variance $\sigma^2$ and $\{H_q(q)\}$ a sequence of filters such that $\sup_{|z| = 1} |H_q(z) - 1| = o_p(1)$ [37, Theorem 3.1]. Therefore, the application of the filter $1/A^{(k+1)}(t)$ yields data $\{u_q^F, y_q^F\}$ such that

$u_q^F = \tilde{u}_q^F + o_p(1)$

$y_q^F = \tilde{y}_q^F + o_p(1)$

where $\{\tilde{u}_q^F, \tilde{y}_q^F\}$ are such that $A_q(t) \tilde{u}_q^F = B_q(t) \tilde{y}_q^F + e_q$.

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

$\Phi_N = \Phi_N^o + o_p(1)$

where $\Phi_N^o$ is the regressor matrix obtained from $\{\tilde{u}_q^F, \tilde{y}_q^F\}$.

Therefore

$\frac{1}{N} \Phi_N^F \Phi_N^F = \frac{1}{N} \Phi_N^o \Phi_N^o + o_p(1)$

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

$\frac{1}{N} \Phi_N^F \Phi_N^F \to \Gamma \succ 0$
for some $\Gamma > 0$. This verifies the second assumption of Section IV-A-1. Finally, notice that

\[
V(\theta, D_N) = \frac{1}{N} \left\{ Y_N^F - \Phi_N^T \Phi_N \theta \right\}_2^2 = \frac{1}{N} \left\{ \theta - \left( \frac{\Phi_N^T \Phi_N}{N} \right)^{-1} \left( \frac{\Phi_N^T \Phi_N}{N} \right)^T Y_N^F \right\}_2^2
\]

\[
+ \frac{1}{N} \left( Y_N^F \right)^T \left[ I - \Phi_N^T \left( \frac{\Phi_N^T \Phi_N}{N} \right)^{-1} \left( \frac{\Phi_N^T \Phi_N}{N} \right)^T \right] Y_N^F
\]

where $Y_N^F = [y_1^F \cdots y_n^F]^T$, hence

\[
V \left( \hat{\theta}_{LS}^N, D_N \right) = \frac{1}{N} \left( Y_N^F \right)^T \left[ I - \Phi_N^T \left( \frac{\Phi_N^T \Phi_N}{N} \right)^{-1} \left( \frac{\Phi_N^T \Phi_N}{N} \right)^T \right] Y_N^F.
\]

Now,

\[
\Phi_N^{(k)} \left( \frac{\Phi_N^T}{N} \right)^{-1} \left( \frac{\Phi_N^T}{N} \right)^T \Phi_N^{(k)} = \frac{1}{N} \varphi_{(k)} + o_p(1)
\]

\[
\left\{ \frac{1}{N} \left( \frac{\varphi_{(k)}^{\ominus}}{N} \right)^\varphi_{(k)} + o_p(1) \right\} \left\{ \varphi_{(k)} + o_p(1) \right\}^T
\]

\[
\left\{ \frac{1}{N} \left( \frac{\varphi_{(k)}^{\ominus}}{N} \right)^\varphi_{(k)} + o_p(1) \right\} \left\{ \varphi_{(k)} + o_p(1) \right\}^T
\]

\[
\Phi_N^{(k)} (\varphi_{(k)}^{\ominus})^{-1} \varphi_{(k)}^{\ominus} + o_p(1)
\]

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

\[
V \left( \hat{\theta}_{LS}^N, D_N \right) = \frac{1}{N} \left( Y_N^F \right)^T \left[ I - \Phi_N^T \left( \frac{\Phi_N^T \Phi_N}{N} \right)^{-1} \left( \frac{\Phi_N^T \Phi_N}{N} \right)^T \right] Y_N^F
\]

\[
= \frac{1}{N} \left( Y_N^F \right)^T \left[ I - \Phi_N^T \left( \frac{\Phi_N^T \Phi_N}{N} \right)^{-1} \left( \frac{\Phi_N^T \Phi_N}{N} \right)^T \right] Y_N^F
\]

using [2, Theorem 2.B.1] again. This verifies the first assumption of Section IV-A-1, 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-A-1. Furthermore, expanding (8) and using that $(I - \Phi_N^T \Phi_N)^{-1} \Phi_N^T$ is idempotent gives

\[
N \cdot V(\theta, D_N) = \left\| \theta - (\Phi_N^T \Phi_N)^{-1} \Phi_N^T Y_N \right\|_{L_N}^2
\]

\[
+ \left\| (I - \Phi_N (\Phi_N^T \Phi_N)^{-1} \Phi_N^T) E_N \right\|_2^2
\]

(40)

For $\hat{\theta}_{LS}^N = (\Phi_N^T \Phi_N)^{-1} \Phi_N^T Y_N$, (40) reveals that

\[
\frac{1}{\sigma^2} V \left( \hat{\theta}_{LS}^N, D_N \right) = \frac{1}{N} \left\| (I - \Phi_N (\Phi_N^T \Phi_N)^{-1} \Phi_N^T) E_N \right\|_{L_N}^2
\]

These two observations imply that

\[
(1 + \varepsilon_N V \left( \hat{\theta}_{LS}^N, D_N \right) \geq V(\theta, D_N)
\]

or, equivalently,

\[
\varepsilon_N V \left( \hat{\theta}_{LS}^N, D_N \right) \geq \left\| \theta - (\theta^o + (\theta^o)^{-1} \Phi_N E_N \right\|_{L_N}^2
\]

Proof of Lemma II.2

If $0 \in \Omega$, then it trivially follows that $\theta = 0$ is the unique optimum of (34). Now let us assume for the rest of the proof that $0 \notin \Omega$. Notice that $\|\cdot\|_1$ is a norm in $\mathbb{R}^n$. Since $\Omega$ is a closed set in the topology of $\|\cdot\|_N$, it is also closed in the topology of $\|\cdot\|_1$ (since all norms in $\mathbb{R}^n$ are topologically equivalent [38, Problem 15.7]). Hence, dist(\Omega, 0) = inf{ $\|w \circ \theta\|_1: \theta \in \Omega$} = $\lambda \geq 0$, and there is an element $\theta^* \in \Omega$ such that $\|\omega \circ \theta^*\|_1 = 1$. Such $\theta^*$ is an optimum of (34).

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

\[
\|\omega \circ (\theta - \bar{\theta})\|_1 = \left\| \omega \circ \left( \left\{ \|\omega \circ \theta\|_1 - \lambda/2 \right\} \theta \right\|_1
\]

\[
= \|\omega \circ \theta\|_1 - \lambda/2 \right\|_1 \|\omega \circ \theta\|_1 = \frac{\lambda}{2}
\]

hence $\bar{\theta} \in U \subset \Omega$, but

\[
\|\omega \circ \bar{\theta}\|_1 = \left\| \omega \circ \left( \left\{ \|\omega \circ \theta\|_1 - \lambda/2 \right\} \bar{\theta} \right\|_1
\]

\[
= \delta - \lambda/2 < \delta.
\]

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

Proof of Lemma III.1

First notice that due to the second and third assumptions of the lemma, $\|\theta - (\theta^0 + \varepsilon)\|_{L_N}^2 \leq \sigma^2 \varepsilon_N \eta$ can hold only if $\theta = [\theta^{(1) \theta^{(2)}}, \ldots, \theta^{(n)}]$. With $\theta^{(i)} \in \mathbb{R}^{n}$ $(i = 1, 2)$ and $\theta^{(1)} \neq 0$ for every $k \in \mathbb{N}$. Otherwise, if $[\theta^{(1)}]_k = 0$ for some $k \in \mathbb{N}$,

\[
[\theta^{(i)}]_k = [\theta^{(i)}]_k - [\theta^{(i)}]_k \leq \frac{\sum_{i=1}^n \theta^{(i)} - (\theta^{(i)})_i + \xi_i)^2}{1 + \lambda_{\min}(I_N)} [\theta - (\theta^{(i)} + \xi)]_{1,1} \leq 0.5 \min_{i \in \mathbb{N}} \left( \left\{ \theta^{(i)} \right\}_{1,1} \right)
\]

which is a contradiction. Now, let us further partition $\theta^{(2)}$ as $\theta^{(2)} = [\theta^{(2)}_1, \theta^{(2)}_2, \ldots, \theta^{(2)}_n]$, where $\theta^{(2)}_k$ corresponds to those $\theta^{(2)}$, which are associated with the parameters of $A$, and partition $w_N$ and $\xi$ accordingly. Note that these parameters are exactly the zero parameters allocated at $A$. $\theta^{(2)}_k$ is defined respectively.
Our goal is to show that this partition leads to a contradiction if $n_a > 0$.

By the assumptions of the lemma, $0 \not\in \Omega$, hence by Lemma II.2, the optimum lies in the boundary of $\Omega$. Therefore, the optimality conditions for problem (34), omitting the complementary conditions, are [39, Section 28]

$$0 \in \partial \theta \left( ||w \circ \theta||_1 + (\mu/2) \left[ \left( \theta - (\theta^0 + \xi) \right) ||w||_{\Omega_N}^2 - \sigma^2 \varepsilon_N \eta \right] \right)$$

$$\mu \geq 0, \quad \left[ \theta - (\theta^0 + \xi) \right] ||w||_{\Omega_N}^2 = \sigma^2 \varepsilon_N \eta \quad (41)$$

for some $\mu$, where $\partial \theta f(\theta)$ denotes the subdifferential of a function with respect to $\theta$. Note that the regularity conditions for (41) to be necessary and sufficient hold, since the constraint set contains an interior point, e.g., $\theta = \theta^0 + \xi$. After some algebra, using facts such as $\partial||x||_1 = \text{Sgn}(x)$, and the partition of $\theta^0$ and $\omega$, we can rewrite (41) as

$$\begin{bmatrix} (\theta_1) - (\theta_0 + 1) \\
\theta_2 \\
\cdots \\
\theta_n 
\end{bmatrix} = -\frac{1}{\mu} \begin{bmatrix} w_n^{(1)} \circ \text{Sgn} (\theta_1) \\
\vdots \\
\vdots \\
w_n^{(n)} \circ \text{Sgn} (\theta_n) 
\end{bmatrix},$$

$$\mu \geq 0, \quad \left[ \theta - (\theta^0 + \xi) \right] ||w||_{\Omega_N}^2 = \sigma^2 \varepsilon_N \eta. \quad (42)$$

We can partition $\Gamma_N^{-1}$ according to the partition of $\theta$ as

$$\Gamma_N^{-1} = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\
M_{21} & M_{22} & M_{23} \\
M_{31} & M_{32} & M_{33} 
\end{bmatrix},$$

which, together with (42), gives

$$\begin{aligned}
\theta_n^{(2)} &= -\frac{1}{\mu} \left[ M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} \right] \left[ w_n^{(a)} \circ \text{Sgn} (\theta_n^{(2)}) \right] \\
&\quad + \xi_n^{(2)} - M_{a,b}M^{-1}_{b,n} \xi_n^{(2)} \\
&\quad - \frac{1}{\mu} \left[ M_{a,1} - M_{a,b}M^{-1}_{b,1}M_{b,1} \right] \left[ w_n^{(1)} \circ \text{Sgn}(\theta_1^{(a)}) \right].
\end{aligned} \quad (43)$$

Note that $M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} > 0$ (since $\Gamma_N > 0$, [34, Theorem 7.7.6]). Therefore, (43) and the assumptions of the lemma imply the sparseness of $\theta_N^0$ by noting that

$$0 < \sum_{i=1}^{n_a} \left[ w_n^{(a)} \right]_i \left[ \theta_n^{(2)} \right]_i \leq q^T \theta_n^{(2)} =$$

$$-\frac{1}{\mu} q^T \left[ M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} \right] q + q^T \left( \xi_n^{(2)} - M_{a,b}M^{-1}_{b,n} \xi_n^{(2)} \right) - \frac{1}{\mu} \left[ M_{a,1} - M_{a,b}M^{-1}_{b,1}M_{b,1} \right] \left[ w_n^{(1)} \circ \text{Sgn}(\theta_1^{(a)}) \right] < 0 \quad (44)$$

where $q = w_n^{(a)} \circ \text{Sgn}(\theta_n^{(2)})$, 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

$$\begin{aligned}
\mu &\geq \frac{1}{\sqrt{\sigma^2 \varepsilon_N \eta}} \frac{1}{\min \left( \Gamma_N^{-1} \right)} \left[ \left[ w_n^{(1)} \right]_2^2 + \left[ w_n^{(2)} \right]_2^2 \right] \\
\mu &\leq \frac{1}{\sqrt{\sigma^2 \varepsilon_N \eta}} \sqrt{\left| \Gamma_N^{-1} \right|} \sqrt{\left[ w_n^{(1)} \right]_2^2 + \left[ w_n^{(2)} \right]_2^2}.
\end{aligned}$$

Based on these inequalities, we have

$$\begin{aligned}
\frac{1}{\mu} q^T \left[ M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} \right] q &\geq \frac{1}{\mu} \lambda_{\min} \left\{ M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} \right\} \left[ w_n^{(2)} \right]_2^2 \\
&\geq \left( \frac{1}{\mu} \lambda_{\min} \left\{ M_{a,n} - M_{a,b}M^{-1}_{b,n}M_{b,a} \right\} \right) \left[ w_n^{(2)} \right]_2^2.
\end{aligned}$$

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