Preprints of the 16th IFAC Symposium on System Identification
The International Federation of Automatic Control
Brussels, Belgium. July 11-13, 2012

Sparse Estimation of Rational Dynamical Models

Roland Tóth * Håkan Hjalmarsson ** Cristian R. Rojas **

* Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands
** Automatic Control Lab and ACCESS Linnaeus Center, Electrical Engineering, KTH–Royal Institute of Technology, S-100 44 Stockholm, Sweden

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. This can be motivated either from appealing to a parsimony principle (*Occam's razor*) or from the view point of the utilization complexity in terms of control synthesis, prediction, etc. 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, which is often implemented as solving an optimization program based on a convex relaxation (e.g. ℓ_1/LASSO, nuclear norm, ...). However, in order to apply these methods, the (unpenalized) cost function must be convex. This imposes a severe constraint on the types of model structures or estimation methods on which these relaxations can be applied. In this paper, we extend the use of convex relaxation techniques for sparsity to general rational plant model structures estimated by using prediction error minimization. This is done by combining the LASSO and the Steiglitz-McBride approaches. To demonstrate the advantages of the proposed solution an extensive simulation study is provided.

1. 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 [Eykhoff, 1974, Ljung, 1999, Söderström and Stoica, 1989, Pintelon and Schoukens, 2001]. It also has a close connection with allied fields such as statistics, econometrics, machine learning and chemometrics [Ljung, 2010].

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. As 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 utilization complexity in terms of control synthesis, prediction, etc. Since an optimal choice in this question is rarely known a priori, an identification user typically proposes a model structure capable to 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 substructure (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 [Weisberg, 1980]. *Forward Stage-wise Selection* and *Least Angle Regression* (LARS) [Efron et al., 2004] 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 [Donoho and Johnstone, 1994]. 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) [Tibshirani, 1996] and the *Non-Negative Garrote* (NNG) [Breiman, 1995], 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).

Most of the aforementioned sparse estimation methods can only be applied to model structures of a linear regression

* This work was supported by the European Research Council under the advanced grant LEARN, contract 267381.

type (*i.e.*, where the cost function to be minimized by the estimator is quadratic). Some extensions, however, have been conceived for estimators based on the minimization of a convex loss function [Bühlmann and van de Geer, 2011, 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 [Bühlmann and van de Geer, 2011, 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 [Ljung, 1999]) 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 [Ljung, 1999]) may end up having multiple local optima if an ℓ_1 regulation term is added to it to impose sparsity.

In this paper, we extend the use of convex relaxation techniques for sparsity to general LTI rational Output-Error-type model structures estimated using *Prediction Error Methods* (PEM), where we allow the noise to be colored. To this end, we first combine a variant of the LASSO called SPARSEVA [Rojas and Hjalmarsson, 2011], and the *Steiglitz-McBride method*, which is a technique for the estimation of *Output-Error* (OE) models. Since the Steiglitz-McBride approach reduces the problem of estimation 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, thus imposing sparsity in the resulting plant model, when 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 [Zhu, 2011].

The paper is organized as follows. Section 2 introduces the problem formulation. A description of the technique proposed is given in Section 3 after a brief description of SPARSEVA and the Steiglitz-McBride methods. Section 5 presents several simulation examples that show the properties of our proposed methods. Finally, the paper is concluded in Section 6.

2. PROBLEM STATEMENT

Consider the stable discrete-time LTI data-generating system

$$y_t = \frac{B_{\mathrm{o}}(q)}{A_{\mathrm{o}}(q)} u_t + \frac{C_{\mathrm{o}}(q)}{D_{\mathrm{o}}(q)} e_t, \tag{1}$$

where $\{e_t\}$ is a Gaussian white noise sequence of zero mean and variance $\sigma_{\mathrm{e}}^2 > 0$, $\{u_t\}$ is a quasi-stationary signal [Ljung, 1999], and

$$\begin{aligned} A_{\mathrm{o}}(q) &= 1 + a_1^{\mathrm{o}} q^{-1} + \cdots + a_{n_{\mathrm{a}}}^{\mathrm{o}} q^{-n_{\mathrm{a}}}, \\ B_{\mathrm{o}}(q) &= b_1^{\mathrm{o}} q^{-1} + \cdots + b_{n_{\mathrm{b}}}^{\mathrm{o}} q^{-n_{\mathrm{b}}}, \\ C_{\mathrm{o}}(q) &= 1 + c_1^{\mathrm{o}} q^{-1} + \cdots + c_{n_{\mathrm{c}}}^{\mathrm{o}} q^{-n_{\mathrm{c}}}, \\ D_{\mathrm{o}}(q) &= 1 + d_1^{\mathrm{o}} q^{-1} + \cdots + d_{n_{\mathrm{d}}}^{\mathrm{o}} q^{-n_{\mathrm{d}}}, \end{aligned}$$

with $\theta_{\mathrm{o}} = [\, a_1^{\mathrm{o}} \ldots a_{n_{\mathrm{a}}}^{\mathrm{o}} \; b_1^{\mathrm{o}} \ldots b_{n_{\mathrm{b}}}^{\mathrm{o}} \,]$ and $\eta_{\mathrm{o}} = [\, c_1^{\mathrm{o}} \ldots c_{n_{\mathrm{c}}}^{\mathrm{o}} \; d_1^{\mathrm{o}} \ldots d_{n_{\mathrm{d}}}^{\mathrm{o}} \,]$. 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. Our goal is to estimate a model of this system based on measurements $\{u_t, y_t\}_{t=1}^N$, of the form

$$y_t = \frac{B(q)}{A(q)} u_t + \frac{C(q)}{D(q)} \epsilon_t, \tag{2}$$

where

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

In this paper we assume that the model structure (2) contains the true system (1), *i.e.*, there is no undermodelling.

3. PROPOSED METHOD

In this section, we propose a method for the estimation of model structure (2) taking into account the sparsity in the parameter vector. To this end, first we present some preliminaries on SPARSEVA (a sparse LASSO-type estimator) and the Steiglitz-McBride method. Later, we show how to combine these two procedures in order to estimate general sparse rational plant model structures.

3.1 SPARSEVA

To introduce this method, developed in [Rojas and Hjalmarsson, 2011], let us restrict the model (2) to an equation error structure with $C(q) = 1$ and $D(q) = A(q)$, *i.e.*,

$$A(q) y_t = B(q) u_t + \epsilon_t. \tag{3}$$

This model can be written in a linear regression fashion as

$$Y_N = \Phi_N \theta + E_N,$$

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

$$\Phi_N = \begin{bmatrix} -y_{n_{\mathrm{a}}} & \cdots & -y_1 & u_{n_{\mathrm{a}}} & \cdots & u_{n_{\mathrm{a}}-n_{\mathrm{b}}+1} \\ \vdots & & \vdots & \vdots & & \vdots \\ -y_{N-1} & \cdots & -y_{N-n_{\mathrm{a}}} & u_{N-1} & \cdots & u_{N-n_{\mathrm{b}}} \end{bmatrix}. \tag{4}$$

(For simplicity of presentation, we assume that $n_a \geq n_b$.) The SPARSEVA estimator (which stands for *SPARSe Estimator based on a VAlidation criterion*) is a variant of the LASSO estimator [Tibshirani, 1996] (an ℓ_1 penalized least-squares estimator), and it corresponds to the minimizer of the convex program:

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|_1 \\ \text{s.t.} \quad & V_N(\theta) \leq V_N(\hat{\theta}_N^{\mathrm{LS}})(1 + \varepsilon_N). \end{aligned} \tag{5}$$

Here $\hat{\theta}_N^{\mathrm{LS}} := (\Phi_N^\top \Phi_N)^{-1} \Phi_N^\top Y_N$ is the least-squares estimator of θ, $V_N(\theta) := \frac{1}{N} \|Y_N - \Phi_N \theta\|_2^2$ is the least-squares cost function, and $\varepsilon_N > 0$ is a quantity which can be typically chosen as:

- $\varepsilon_N = 2(n_{\mathrm{a}} + n_{\mathrm{b}})/N$.
- $\varepsilon_N = (n_{\mathrm{a}} + n_{\mathrm{b}}) \ln(N)/N$.

The first choice is motivated by the AIC criterion, while the second one is related to the BIC/MDL criterion [Ljung, 1999]. In Section 4, it is going to be shown how these choices of ε_N relate to consistency/sparsity of the SPARSEVA scheme.

Even though SPARSEVA can be considered as a variant of the LASSO, it has the advantage of not requiring the tuning of regularization parameters via techniques such as cross-validation, which involve solving multiple times a convex program over a grid of values of the regularization parameters. This tuning is automatically taken into account by choosing the value of ε_N, as explained in detail in [Rojas and Hjalmarsson, 2011].

An "adaptive" version of (5), called A-SPARSEVA, has better sparsity properties than SPARSEVA [Rojas and Hjalmarsson, 2011], and is defined as the minimizer of the following convex program:

$$\begin{aligned} \min_{\theta} \quad & \|W\theta\|_1 \\ \text{s.t.} \quad & V_N(\theta) \le V_N(\hat{\theta}_N^{\mathrm{LS}})(1+\varepsilon_N), \end{aligned} \tag{6}$$

where $W := \mathrm{Diag}([\hat{\theta}_N^{\mathrm{LS}}]_1^{-1}, \dots, [\hat{\theta}_N^{\mathrm{LS}}]_{n_\mathrm{a}+n_\mathrm{b}}^{-1})$.

The estimation properties of A-SPARSEVA can be further improved by removing the columns of Φ corresponding to the zero entries of the A-SPARSEVA estimate $\hat{\theta}^{\mathrm{A}}$, and re-estimating θ by least-squares on the reduced Φ.

The properties of SPARSEVA and its variants are discussed in Section 4.

3.2 Steiglitz-McBride Method

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

$$y_t = \frac{B(q)}{A(q)} u_t + \epsilon_t, \tag{7}$$

which corresponds to (2) with $C(q) = D(q) = 1$. It is well known [Ljung, 1999] 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 (4)) is biased, and the cost function of the *Prediction Error Method* (PEM) [Ljung, 1999] 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 (7) from least-squares fits is the so-called Steiglitz-McBride method [Steiglitz and McBride, 1965]. The idea of this method is to iteratively prefilter u_t and y_t by $1/\hat{A}^{(k)}(q)$, where $\hat{A}^{(k)}(q)$ is an estimate of the $A(q)$ polynomial (at step k), and then to apply least-squares to the data, assuming a model structure such as (3), 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 [Stoica and Söderström, 1981, Regalia, 1995]. In particular, it is known to give unbiased estimates only if the true system belongs to an OE structure (7), and its global convergence properties are still largely an open problem. In addition, the Steiglitz-McBride is not asymptotically efficient for (7).

In [Zhu, 2011], an interesting extension of the Steiglitz-McBride algorithm has been developed, which gives consistent estimates even for Box-Jenkins model structures (2). This extension consists in performing a preliminary step, where a high order ARX model

$$A_{\mathrm{HO}}(q) y_t = B_{\mathrm{HO}}(q) u_t + \epsilon_t, \tag{8}$$

with

$$\begin{aligned} A_{\mathrm{HO}}(q) &= 1 + a_1^{\mathrm{HO}} q^{-1} + \cdots + a_m^{\mathrm{HO}} q^{-m}, \\ B_{\mathrm{HO}}(q) &= b_1^{\mathrm{HO}} q^{-1} + \cdots + b_m^{\mathrm{HO}} q^{-m}, \end{aligned}$$

is fitted to the data, and used then to prefilter the data, *i.e.*, to generate the signals

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

This filtered data is then used in place of the original input and output signals of (7) 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 prefiltering 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 $B(q)$.

Some results on the accuracy of the extended Steiglitz-McBride method are detailed in Section 4.

3.3 Estimation of Sparse Output-Error Models

As mentioned in Section 3.1, SPARSEVA and ℓ_1-penalized estimators cannot be directly applied to model structures such as (2), 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.

Algorithm 1 OE-SPARSEVA with Steiglitz-McBride

```
Require: a data record $\mathcal{D}_N = \{u_t, y_t\}_{t=1}^N$ of (1) and the
    model structure (7) 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. (7) and (7) is globally identifiable
    on $\Theta$ [Ljung, 1999].
1: Let $m \gg n_\mathrm{a}$ and fit using least-squares the high order
    ARX model described by (8) 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
           $A^{(k)}(q) y_t^{\mathrm{F}(k)} = B^{(k)}(q) u_t^{\mathrm{F}(k)} + \epsilon_t,$
       resulting in the estimates $\hat{A}^{(k)}$, $\hat{B}^{(k)}$ and the associ-
       ated 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
       $A(q) y_t^{\mathrm{F}(k+1)} = B(q) u_t^{\mathrm{F}(k+1)} + \epsilon_t.$
9: return estimated model (7).
```

Based on the previous discussion, Algorithm 1 provides estimation of sparse rational OE models (7).

Remark 1. 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 (6) can be modified so that only the coefficients of $A(q)$ are included.

Remark 2. 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 $V_N(\hat{\theta}_N^{\mathrm{LS}})$ in (5) is not necessary). This might refine the results for relatively small data-lengths N under considerable noise, but at the expanse 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.

4. THEORETICAL RESULTS

In this section, some theoretical support for the method proposed in Section 3 is provided.

To start, A-SPARSEVA enjoys the properties presented in the following theorem.

Theorem 3. Under the assumptions of Sections 2 and 3.1:

(1) The A-SPARSEVA estimator $\hat{\theta}_N$ is consistent in probability if and only if $\varepsilon_N \to 0$ as $N \to \infty$.
(2) Under the condition for consistency in probability (i.e., $\varepsilon_N \to 0$ as $N \to \infty$, c.f. statement (1)), $\hat{\theta}_N$ has the *sparseness property* (*i.e.*, $P\{(\hat{\theta}_N)_i = 0\} \to 1$ as $N \to \infty$ for every index i such that $[\theta_{\mathrm{o}}]_i = 0$) if and only if $N\varepsilon_N \to \infty$.
(3) If $\varepsilon_N \to 0$, but $N\varepsilon_N \to \infty$ as $N \to \infty$, then $\hat{\theta}_N$ (with least-squares re-estimation) has the *oracle property*. This means that
$$\sqrt{N}(\hat{\theta}_N - \theta_{\mathrm{o}}) \in \mathrm{As}\,\mathcal{N}(0, M^{\dagger}),$$
where M is the information matrix when the support of θ_{o} is known (and it is such that $M_{ik} = 0$ whenever $(\theta_{\mathrm{o}})_i = 0$ or $(\theta_{\mathrm{o}})_k = 0$).

This theorem essentially corresponds to Theorems 3.1, 3.2 and 3.3 of [Rojas and Hjalmarsson, 2011]. The proofs in [Rojas and Hjalmarsson, 2011] apply to the case when the regressor matrix Φ_N is deterministic. However, these proofs can be easily extended to the case considered in Section 3.1, by noting that they apply if the following two properties hold:

(1) $V_N(\hat{\theta}_N^{\mathrm{LS}}) \to \sigma_{\mathrm{e}}^2$ in probability as $N \to \infty$, and
(2) $\sqrt{N}(\hat{\theta}_N^{\mathrm{LS}} - \theta_{\mathrm{o}}) \in \mathrm{As}\,\mathcal{N}(0, \sigma_{\mathrm{e}}^2 M)$, where M is a non-singular matrix.

These properties hold under the assumptions of Sections 2 and 3.1 (see, e.g., [Ljung, 1999]), hence implying the validity of Theorem 3. For the non-asymptotic properties of the method, especially in case of relatively small data records, see [Tóth et al., 2011].

Notice that, from Theorem 3, A-SPARSEVA enjoys consistency, sparseness and the oracle property if we choose $\varepsilon_N = (n_{\mathrm{a}} + n_{\mathrm{b}})\ln(N)/N$, resembling the BIC criterion.

The modified Steiglitz-McBride method presented in this paper is due to Y. Zhu [Zhu, 2011]. 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., [Stoica and Söderström, 1981]), 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 efficient estimator of $A_{\mathrm{o}}(q)$ and $B_{\mathrm{o}}(q)$ for general Box-Jenkins model structures [1] (2).

The combination of A-SPARSEVA and the modified Steiglitz-McBride method, as presented in Section 3.3, can be expected to have attractive asymptotic properties. In particular, by combining the theoretical results of its components, we believe that this technique is consistent in probability and has the sparseness property if the sequence ε_N is chosen to decay to zero at a rate even slower than with $\ln(N)/N$. This is because A-SPARSEVA is based on data which has been prefiltered by consistent estimates of A, and do not satisfy exactly the model structure (3).

Remark 4. Notice that the scaling of the parameters in θ_{o} does not seem to play a major role in the estimation performance of Algorithm 1, at least asymptotically in N, since A-SPARSEVA weights the ℓ_1 norm by the inverse of the estimates in $\hat{\theta}_N^{\mathrm{LS}}$, which compensates for the relative size of the components of θ_{o}.

5. NUMERICAL EXAMPLES

Consider the data-generating system (1) described by the following polynomials:
$$A_{\mathrm{o}}(q) = -1.42q^{-2} + 0.5q^{-4}, \quad B_{\mathrm{o}}(q) = 1.3 + 1.2q^{-4},$$
$$C_{\mathrm{o}}(q) = 1, \qquad D_{\mathrm{o}}(q) = 1.$$
This system obviously has an OE type of noise structure. To identify this system from data based on the previously proposed estimation scheme, consider the model structure (7) with $n_{\mathrm{a}} = 5$ and $n_{\mathrm{b}} = 5$. Even if this corresponds to a rather accurate guess of the original order of the polynomials involved, the true parameter vector
$$\theta_{\mathrm{o}} = [\,0 \;\; -1.42 \;\; 0 \;\; 0.5 \;\; 0 \;\; 1.3 \;\; 0 \;\; 0 \;\; 0 \;\; 1.2\,]$$
corresponding to the data-generating system is rather sparse.

For estimation purposes, 100 estimation and 100 validation data records have been 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 $u_t \in \mathcal{N}(0, 1)$ and $e_t \in \mathcal{N}(0, \sigma_{\mathrm{e}}^2)$ respectively. To study the effect of a change in the power of the noise, this generation of the data sequences have been repeated for various noise variances $\sigma_{\mathrm{e}}^2 \in \{0.01, 0.5, 2, 4\}$ corresponding to average *Signal to Noise Ratios* [2] (SNR's): 118dB, 48dB, 25dB, 15dB 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 1 with $m = 50$ and with LS re-optimization 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

[1] Even though it is possible to propose variants of Algorithm 1, 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 are convergent. This suggests that not much may be gained by considering other variants of Algorithm 1.

[2] 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 $v_t = \frac{C_{\mathrm{o}}(q)}{D_{\mathrm{o}}(q)} e_t$.

of the estimates, a base-line estimator or so called `oracle` estimator in terms of the Steiglitz-McBride method has been also applied with the priori knowledge of which elements of θ_o is zero. Note that the latter approach cannot be applied in practice as the optimal model structure is unknown (part of the identification problem itself). The results are compared in terms of

- The *Mean Squared Error* (MSE) of the prediction on the validation data:
$$\text{MSE} = \frac{1}{N}\mathbb{E}\{\|y(k) - \hat{y}_{\hat{\theta}_N}(k)\|_2^2\}.$$
computed as an average over each 100 runs for a given N and σ_e^2.
- The average of the *fit score* or the *Best Fit Rate* (BFR) [Ljung, 2006]:
$$\text{BFR} = 100\% \cdot \max\left(1 - \frac{\|y(k) - \hat{y}_{\hat{\theta}_N}(k)\|_2}{\|y(k) - \bar{y}\|_2}, 0\right),$$
where $\bar{y}$ is the mean of y and $\hat{y}_{\hat{\theta}}$ is the simulated model output based on the validation data.
- The ℓ_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 Figure 1 and the mean and standard deviation of the parameters are given in the SNR= 25dB, $N = 2000$ case in Table 1. From these results it follows that in the low noise cases (SNR=118dB, 48dB) the proposed OE-SPARSEVA scheme correctly estimates the true support of θ_o, *i.e.*, it correctly identifies the underlying model structure of the system and hence it achieves the same results as the `oracle` approach. The performance difference of the `oe` approach and the `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=25dB), for small data lengths we can observe that OE-SPARSEVA loses the benefits of the regularized optimization scheme by overestimating 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 better results than the `oe` method. We can see that the performance of OE-SPARSEVA asymptotically converges to the `oracle` approach while the `oe` has a much slower convergence rate. The same behavior can be observed in the SNR=15dB case where the “point of recovery” is around $N = 1600$.

6. CONCLUSIONS

It has been shown that by combining the SPARSEVA approach 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 structures in which the underlying data-generating system is represented by a sparse parameter vector. A main benefit of the method 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 simulations based Monte-Carlo study.

REFERENCES

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

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

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–451, 2004.

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

L. Ljung. *System Identification: Theory for the User, 2nd Edition.* Prentice Hall, 1999.

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

L. Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34:1–12, 2010.

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

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

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.

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

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

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

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

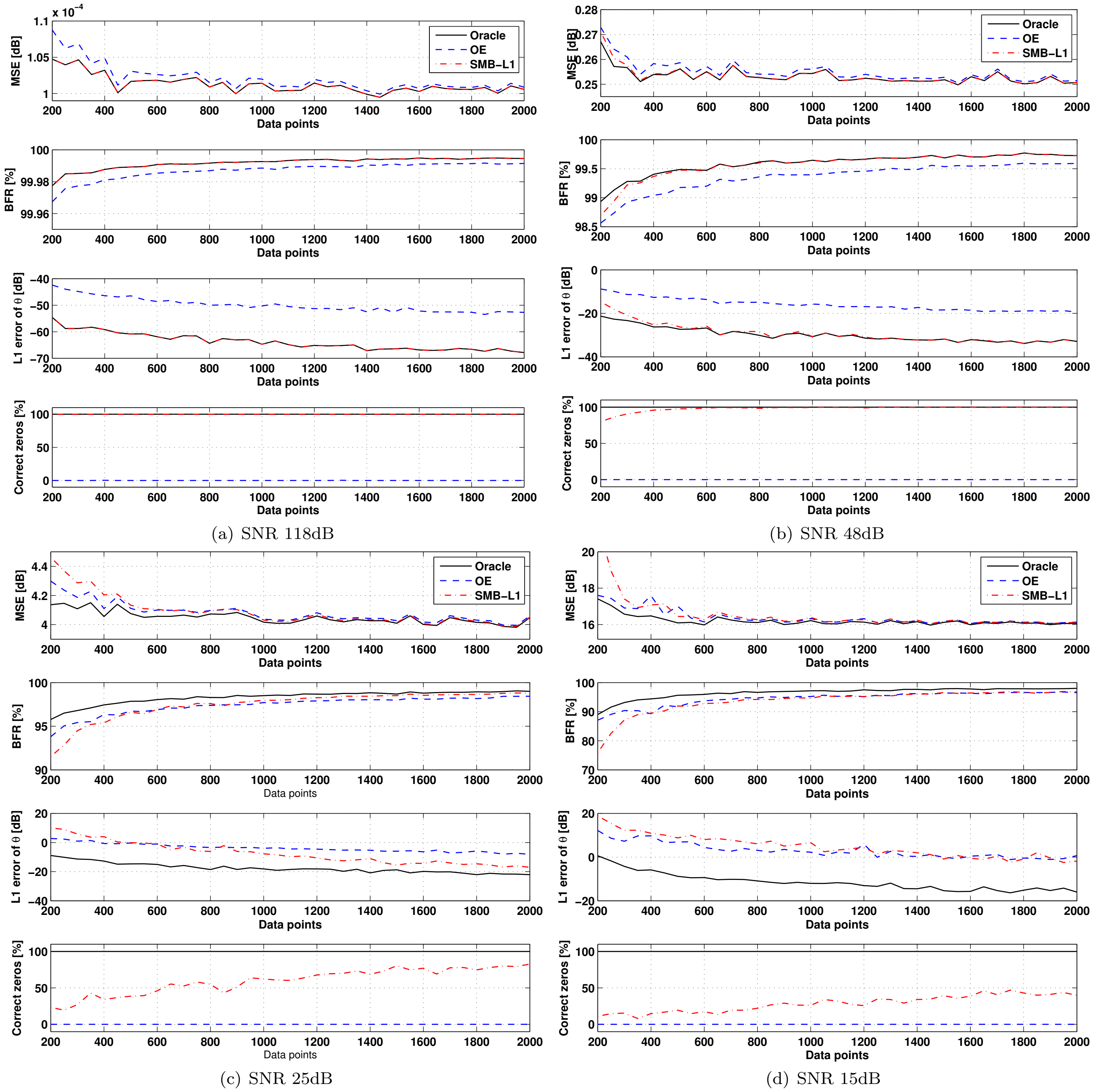
R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent. Compressive system identification in the linear time-invariant framework. In *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011.

S. Weisberg. *Applied Linear Regression.* Wiley, 1980.

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

Table 1. Bias and variance results of the parameter estimates by the `oracle`, `oe` and the OE-SPARSEVA methods in the SNR= 25dB, $N = 2000$ case.

Method		a_1	a_2	a_3	a_4	a_5	b_0	b_1	b_2	b_3	b_4
θ_0		0	-1.4200	0	0.5000	0	1.3000	0	0	0	1.2000
`oracle`	mean	0	-1.4199	0	0.4999	0	1.2982	0	0	0	1.2006
	std	0	0.0115	0	0.0102	0	0.0283	0	0	0	0.0512
`oe`	mean	0.0216	-1.4207	-0.0334	0.5006	0.0132	1.2992	0.0198	-0.0050	0.0217	1.1977
	std	0.0533	0.0115	0.0832	0.0102	0.0336	0.0517	0.0731	0.0816	0.0756	0.0676
OE-SPARSEVA	mean	0.0002	-1.4198	-0.0002	0.4999	0	1.2984	-0.0004	0.0016	0.0004	1.2010
	std	0.0039	0.0114	0.0033	0.0102	0	0.0487	0.0097	0.0073	0.0181	0.0654



(a) SNR 118dB

(b) SNR 48dB

(c) SNR 25dB

(d) SNR 15dB

Fig. 1. Monte Carlo simulation results with various SNR's and data lengths N.