Sparse estimation of polynomial dynamical models

Toth, R.; Hjalmarsson, H.; Rojas, C.R.

Published in:
Proceeding of the 16th IFAC Symposium on System Identification, 11-13 July 2012, Brussels, Belgium

DOI:
10.3182/20120711-3-BE-2027.00288

Published: 01/01/2012

Document Version
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us (openaccess@tue.nl) providing details. We will immediately remove access to the work pending the investigation of your claim.

Download date: 24. Jan. 2019
Sparse Estimation of Rational Dynamical Models

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

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 [Eykoff, 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 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 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...
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 $\ell_1$ 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

$$y_t = \frac{B_o(q)}{A_o(q)} u_t + \frac{C_o(q)}{D_o(q)} e_t,$$

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

$$A_o(q) = 1 + a_1 q^{-1} + \cdots + a_n q^{-n},$$
$$B_o(q) = b_1 q^{-1} + \cdots + b_n q^{-n},$$
$$C_o(q) = 1 + c_1 q^{-1} + \cdots + c_m q^{-m},$$
$$D_o(q) = 1 + d_1 q^{-1} + \cdots + d_n q^{-n},$$

with $\theta_o = [a_1 \ldots a_n b_1 \ldots b_n]$, and $\theta_o = [c_1 \ldots c_m d_1 \ldots d_n]$. 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 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 $\{y_t, y_{t-1}\}_{t=1}^n$ of the form

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

where

$$A(q) = 1 + a_1 q^{-1} + \cdots + a_n q^{-n},$$
$$B(q) = b_1 q^{-1} + \cdots + b_n q^{-n},$$
$$C(q) = 1 + c_1 q^{-1} + \cdots + c_n q^{-n},$$
$$D(q) = 1 + d_1 q^{-1} + \cdots + d_n q^{-n}.$$  

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.,

$$A(q)y_t = B(q)u_t + e_t.$$  

This model can be written in a linear regression fashion as

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

where $Y_N := [y_{t+1} \ldots y_N]^\top$, $E_N := [\varepsilon_{t+1} \ldots \varepsilon_N]^\top$, $\theta := [a_1 \ldots a_n b_1 \ldots b_n]^\top$ and

$$\Phi_N := \begin{bmatrix} -y_{t+1} & \ldots & -y_{1} & u_{n+1} & \ldots & u_{n-m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -y_{n-1} & \ldots & -y_{n-m} & u_{n-1} & \ldots & u_{n-m} \end{bmatrix}.$$  

(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 $\ell_1$ penalized least-squares estimator), and it corresponds to the minimizer of the convex program:

$$\min_{\theta} \| \theta \|_1$$

s.t.

$$V_N(\theta) \leq V_N(\hat{\theta}_N^{LS})(1 + \varepsilon_N).$$  

Here

$$\hat{\theta}_N^{LS} := (\Phi_N^\top \Phi_N)^{-1} \Phi_N^\top Y_N$$

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

- $\varepsilon_N = 2(n_a + n_b)/N$.
- $\varepsilon_N = (n_a + n_b) \ln(N)/N$.

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 $\varepsilon_N$ 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 $\varepsilon_N$, 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:

$$
\min_{\hat{\theta}} \|W\theta\|_1
\text{ s.t. } V_N(\hat{\theta}) \leq V_N(\hat{\theta}_{\text{LS}}^N(1+\varepsilon_N),
$$

where $W := \text{Diag}([\hat{\theta}_{\text{LS}}^N]^{-1}, \ldots, [\hat{\theta}]^{-1}_{\text{LS}}_{n_a+n_b})$.

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

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,

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

which corresponds to (2) with $C(q) = D(q) = 1$. It is well known [Ljung, 1999] that the least-squares estimator $\hat{\theta}_{\text{LS}}^N := (\Phi^T \Phi)^{-1} \Phi^T y_N$ (where $\Phi$ 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, 1960]. The idea of this method is to iteratively prefilter $u_t$ and $y_t$ by $1/A(q)$, where $1/A(q)$ is an estimate of the A 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

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

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

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

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 $1/A_{HO}(q)$ should be a reasonable estimate of the noise model $C(m)/D(m)$, 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 $\ell_1$-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 $\ell_1$-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 $\theta = [a_1 \cdots b_m]^T \in \Theta \subseteq \mathbb{R}^{n_a+n_b}$. Assume that $\mathcal{D}_N$ is informative w.r.t. (7) and (7) is globally identifiable on $\Theta$ [Ljung, 1999].

1. Let $n > n_a$ and fit using least-squares the high order ARX model described by (8) to the measurements $\mathcal{D}_N$, resulting in $A_{HO}(q)$ and $B_{HO}(q)$.

2. Filter the data $\mathcal{D}_N$ as

$$
y_t^F := A_{HO}(q)y_t, \quad u_t^F := A_{HO}(q)u_t.
$$

3. Set $k = 0$, and let $\hat{A}^{(0)}(q) = 1, \hat{B}^{(0)}(q) = 0$ and consequently $\hat{\theta}^{(0)} = 0$.

4. **repeat**

5. $k \leftarrow k + 1$ and filter the data $\mathcal{D}_N^F = \{y_t^F, u_t^F\}_{t=1}^N$ as

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

6. Fit using least-squares a model of the form

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

resulting in the estimates $\hat{A}^{(k)}, \hat{B}^{(k)}$ and the associated parameter vector $\hat{\theta}^{(k)}$.

7. **until** $\hat{\theta}^{(k)}$ has converged or the maximum number of iterations is reached.

8. **Apply** A-SPARSEVA (with least-squares re-estimation) to the model

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

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 $\ell_1$-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 $\varepsilon_N$ 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 $V_N(\hat{\theta}_N^S)$ in (5) is not necessary). This might refine the results for relatively small data-lengths $N$ under considerable noise, but at the expense of a much higher computational load. Hence a clearly important feature of the proposed SPARSEVA scheme is an automatic choice of $\varepsilon_N$ 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., $P(\{\hat{\theta}_N^o \neq 0\}) \to 0$) as $N \to \infty$ for every index $i$ such that $[\theta_o]_i = 0$ if and only if $N^2 \varepsilon_N \to \infty$.

3. If $\varepsilon_N \to 0$, but $N^2 \varepsilon_N \to \infty$ as $N \to \infty$, then $\hat{\theta}_N$ (with least-squares re-estimation) has the oracle property.

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 $\Phi$ 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. $V_N(\hat{\theta}_N) \to \sigma^2$ in probability as $N \to \infty$, and

2. $\sqrt{N}(\hat{\theta}_N - \theta_o) \in \mathcal{N}(0, \sigma^2 M)$, 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 $\varepsilon_N = (n_a + n_b) \ln(N)/N$, 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 $A(q)$ and $B(q)$ for general Box-Jenkins model structures. 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 $\varepsilon_N$ 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.

**Remark 4.** Notice that the scaling of the parameters in $\theta_o$ 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 $\ell_1$ norm by the inverse of the estimates in $\hat{\theta}_N^S$, which compensates for the relative size of the components of $\theta_o$.

5. NUMERICAL EXAMPLES

Consider the data-generating system (1) described by the following polynomials:

$$A(q) = -1.42q^{-2} + 0.5q^{-4}, \quad B(q) = 1.3 + 1.2q^{-4}, \quad C(q) = 1, \quad D(q) = 1.$$.

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 $n_a = 5$ and $n_b = 5$. Even if this corresponds to a rather accurate guess of the original order of the polynomials involved, the true parameter vector $\theta_o = [0 -1.42 0 0.5 0 1.3 0 0 0 1.2]$ 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 $N \in \{ 200 + 50k \}^{37}_{k=1}$, 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 $u_i \in \mathcal{N}(0,1)$ and $e_i \in \mathcal{N}(0, \sigma^2_e)$ respectively. To study the effect of a change in the power of the noise, this generation of the data sequences has been repeated for various noise variances $\sigma^2_e \in \{ 0.01, 0.5, 2, 4 \}$ corresponding to average Signal to Noise Ratios $2 \times (SNR's): 118$dB, 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 $SNR := 10 \cdot \log_{10} \left( \frac{\sigma^2_u}{\sigma^2_e} \right)$, where $\sigma^2_u = C_{00}(0) e_i$.

986
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 $\theta_o$ 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:
  \[
  \text{MSE} = \frac{1}{N} \mathbb{E}\{\|y(k) - \hat{y}_{\theta_o}(k)\|_2^2\}.
  \]
  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) [Ljung, 2006]:
  \[
  \text{BFR} = 100\% \cdot \max \left( 1 - \frac{\|y(k) - \hat{y}_{\theta_0}(k)\|_2}{\|y(k) - \bar{y}\|_2}, 0 \right),
  \]
  where $\bar{y}$ is the mean of $y$ and $\hat{y}_{\theta_0}$ is the simulated model output based on the validation data.

- The $\ell_1$ parameter estimation error: $||\theta - \theta_0||_1$.

- 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 parameters of the 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 is that the regularization parameter (or tuning quantity) is automatically chosen, not requiring cross-validation. The derived approach can be used to recover the dynamical structure of the system, i.e., for model structure selection, even in case of heavy over-parametrization or colored noise settings provided that a sufficiently large data set is available. The latter has been demonstrated by an extensive simulations based Monte-Carlo study.

REFERENCES


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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle mean</td>
<td>0</td>
<td>-1.4209</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.2994</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.2994</td>
</tr>
<tr>
<td>oracle std</td>
<td>0</td>
<td>0.0115</td>
<td>0</td>
<td>0.0002</td>
<td>0</td>
<td>0.0284</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0517</td>
</tr>
<tr>
<td>oe mean</td>
<td>0.0216</td>
<td>-1.4207</td>
<td>-0.0334</td>
<td>0.5006</td>
<td>0.0132</td>
<td>1.2992</td>
<td>0.0198</td>
<td>-0.0050</td>
<td>0.0217</td>
<td>1.1977</td>
</tr>
<tr>
<td>oe std</td>
<td>0.0533</td>
<td>0.0114</td>
<td>0.0832</td>
<td>0.0102</td>
<td>0.0336</td>
<td>0.0517</td>
<td>0.0731</td>
<td>0.0816</td>
<td>0.0756</td>
<td>0.0676</td>
</tr>
<tr>
<td>OE-SPARSEVA mean</td>
<td>0.0002</td>
<td>-1.4198</td>
<td>-0.0002</td>
<td>0.4999</td>
<td>0</td>
<td>1.2984</td>
<td>-0.0004</td>
<td>0.0016</td>
<td>0.0004</td>
<td>1.2010</td>
</tr>
<tr>
<td>OE-SPARSEVA std</td>
<td>0.0039</td>
<td>0.0114</td>
<td>0.0033</td>
<td>0.0102</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

988