Toward Tractable Global Solutions to Maximum-Likelihood Estimation Problems via Sparse Sum-of-Squares Relaxations*

Diogo Rodrigues, Mohamed R. Abdalmoaty, and Håkan Hjalmarsson†

***Abstract*— In system identification, the maximum-likelihood method is typically used for parameter estimation owing to a number of optimal statistical properties. However, in many cases, the likelihood function is nonconvex. The solutions are usually obtained by local numerical optimization algorithms that require good initialization and cannot guarantee global optimality. This paper proposes a computationally tractable method that computes the maximum-likelihood parameter estimates with posterior certification of global optimality via the concept of sum-of-squares polynomials and sparse semidefinite relaxations. It is shown that the method can be applied to certain classes of discrete-time linear models. This is achieved by taking advantage of the rational structure of these models and the sparsity in the maximum-likelihood parameter estimation problem. The method is illustrated on a simulation model of a resonant mechanical system where standard methods struggle.**

I. INTRODUCTION

The maximum-likelihood (ML) method is widely used for parameter estimation in several contexts, including system identification, mainly owing to a number of optimal properties [1]–[4]. However, for many commonly used model structures, its computational implementation is complicated by the fact that the resulting optimization problems are nonconvex and the conditions to prove uniqueness of the ML estimator are difficult to verify in practice [5]. Since local optimization algorithms are typically used, their performance largely depends on the choice of the initial estimate and convergence to the global solution cannot be guaranteed.

Several methods have been recently proposed in the system identification literature to alleviate the initialization problem. For example, [6], [7] considered a certain class of linear state-space models and proposed algorithms based on difference of convex programming problems, which may be approximately solved using sequentially convex relaxation. On the other hand, methods based on non-parametric approximations and iterative weighted least-square algorithms have been recently proposed [8], [9]. These methods can be applied to rational linear models, and local optima have been avoided in extensive simulation studies. However, while consistency and asymptotic efficiency can be guaranteed, finite sample properties may differ from ML estimation.

An alternative possibility is to solve the ML parameter estimation problem via appropriate global optimization techniques. Several approaches for global optimization have been reported in the literature. A popular method is the branch-and-bound approach, which divides the space of decision variables into several subsets and seeks the global optimum by estimating upper and lower bounds of the cost and constraints for each subset. Then, the subsets where an optimum cannot be located are excluded from the search and the remaining subsets are subdivided until the global optimum is found [10]. However, being a tree-based method that relies on discretization, it can be computationally expensive. In the worst case, the computational effort grows exponentially with the problem size [11]. Alternative methods based on numerical linear algebra have been suggested [12], [13]. However, these methods have only been applied to very short data sets with a small number of unknowns.

In this paper, another approach is used, namely, the reformulation as a convex problem via the concept of sum-of-squares polynomials, which has been extensively studied in algebraic geometry [14], [15] and applied to a wide range of problems, including applications in control theory, experimental design, parameter initialization, and set-membership estimation [16]–[23]. In particular, it is shown in this paper that the rational structure of certain models, such as the models of discrete-time linear systems, leads to ML parameter estimation problems with sparsity patterns. This fact can then be exploited for a tractable computation of the global solutions via the concept of sum-of-squares polynomials and sparse semidefinite relaxations.

II. PRELIMINARIES

A. ML parameter estimation for linear models

Suppose that we would like to estimate the parameters of a discrete-time, linear time-invariant (LTI), single-input single-output (SISO) model with output $y(t)$ and input $u(t)$. The case of a strictly causal model of order n_x is described by the transfer function

$$P_d(z) = \frac{\sum_{k=1}^{n_x} b_k z^{-k}}{\sum_{k=0}^{n_x} a_k z^{-k}}, \qquad a_0 = 1. \tag{1}$$

Assume that the output $y(t)$ is corrupted by additive noise $e(t)$, such that it is given by the output-error (OE) model

$$y(t) = P_d(z)u(t) + e(t). \tag{2}$$

Then, the inputs and outputs satisfy the relation

$$\sum_{k=0}^{n_x} a_k \left(y(t-k) - e(t-k)\right) = \sum_{k=1}^{n_x} b_k u(t-k),\ t = 1,\ldots,N, \tag{3}$$

with $e(1-n_x) = \ldots = e(0) = u(1-n_x) = \ldots = u(0) = 0$ and $a_0 = 1$, where N is the sample size.

* This work was supported by the VINNOVA Competence Centre AdBIOPRO, contract 2016-05181.

† Diogo Rodrigues, Mohamed R. Abdalmoaty, and Håkan Hjalmarsson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden {diogor,abda,hjalmars}@kth.se

The goal is to estimate the OE model (3) specified by the parameters $\boldsymbol{\theta} := (a_1,\ldots,a_{n_x},b_1,\ldots,b_{n_x})$ and initial conditions $\mathbf{y}_0 := (y(1-n_x),\ldots,y(0))$ based on the observations $\mathbf{y}^N := (y(1),\ldots,y(N))$ and the knowledge of the inputs. In this paper, $e(t)$ denote the output errors that satisfy the system of equations (3) for the true values of the parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{y}_0$, while $\hat{e}(t|\boldsymbol{\theta},\mathbf{y}_0)$ denote the prediction errors that satisfy the same system of equations for arbitrary values of the parameters and initial conditions.

A common method for estimating the parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{y}_0$ from the observations $\mathbf{y}^N$ is the ML method [3], [24], [25]. The ML estimate is defined as the global maximizer of the likelihood function or, equivalently, the log-likelihood function

$$\log L(\boldsymbol{\theta},\mathbf{y}_0,\mathbf{y}^N) := \log p_{\mathbf{Y}^N}(\mathbf{y}^N|\boldsymbol{\theta},\mathbf{y}_0), \tag{4}$$

where $p_{\mathbf{Y}^N}$ denotes the probability density function of the random variable $\mathbf{Y}^N$ from which the sample $\mathbf{y}^N$ is drawn.

The log-likelihood function depends on the probability distribution of $\mathbf{Y}^N$, which is in turn determined by the probability distributions of the random variables $E_1,\ldots,E_N$ with the realizations $e(1),\ldots,e(N)$. If we assume that (i) these random variables are independent and identically distributed (i.i.d.), which implies that $e(1),\ldots,e(N)$ can be seen as realizations of the same random variable E, and (ii) E is a normally distributed random variable with zero mean and variance σ^2, the log-likelihood function becomes

$$\log L(\boldsymbol{\theta},\mathbf{y}_0,\mathbf{y}^N) = -\frac{N\log(2\pi\sigma^2)}{2} - \sum_{t=1}^{N}\frac{\hat{e}(t|\boldsymbol{\theta},\mathbf{y}_0)^2}{2\sigma^2}. \tag{5}$$

The maximization of the log-likelihood function with respect to $\boldsymbol{\theta}$ and $\mathbf{y}_0$ is equivalent to the minimization of the mean squared error (MSE) $\hat{J}(\boldsymbol{\theta},\mathbf{y}_0) := \sum_{t=1}^{N}\hat{e}(t|\boldsymbol{\theta},\mathbf{y}_0)^2/N$. Note that $\hat{J}(\boldsymbol{\theta},\mathbf{y}_0)$ coincides with the cost function of the prediction error method (PEM) with a quadratic loss function [3]. For the sake of simplicity, the prediction error $\hat{e}(t|\boldsymbol{\theta},\mathbf{y}_0)$ is denoted as $\hat{e}(t)$ and the definition $\hat{\mathbf{e}} := (\hat{e}(1),\ldots,\hat{e}(N))$ is used in the remainder of this paper.

Hence, the ML estimation of the parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{y}_0$ in the case of i.i.d. Gaussian noise is formulated as the following constrained optimization problem:

$$\min_{\boldsymbol{\theta},\mathbf{y}_0,\hat{\mathbf{e}}} \sum_{t=1}^{N}\hat{e}(t)^2/N, \tag{6a}$$

$$\text{s.t. } \hat{e}(t) - y(t) + \sum_{k=1}^{n_x} a_k\left(\hat{e}(t-k) - y(t-k)\right) + b_k u(t-k) = 0, \quad t = 1,\ldots,N, \tag{6b}$$

with $\hat{e}(1-n_x) = \ldots = \hat{e}(0) = u(1-n_x) = \ldots = u(0) = 0$. Note that (i) the cost function of this problem can be expressed as $\hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}}/N$, (ii) the first n_x equality constraints can be expressed linearly in the n_x variables $\mathbf{y}_0$ as $\mathbf{A}_{\mathbf{y}_0}(\boldsymbol{\theta},\hat{\mathbf{e}})\mathbf{y}_0 = \mathbf{b}_{\mathbf{y}_0}(\boldsymbol{\theta},\hat{\mathbf{e}})$, and (iii) the N equality constraints can be expressed linearly in the N variables $\hat{\mathbf{e}}$ as $\mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)\hat{\mathbf{e}} = \mathbf{b}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)$.

This means that one can first solve the first n_x equality constraints in the constrained problem (6) for the n_x decision variables $\mathbf{y}_0 = \mathbf{A}_{\mathbf{y}_0}(\boldsymbol{\theta},\hat{\mathbf{e}})^{-1}\mathbf{b}_{\mathbf{y}_0}(\boldsymbol{\theta},\hat{\mathbf{e}})$ and then formulate an equivalent constrained optimization problem as follows:

$$\min_{\boldsymbol{\theta},\hat{\mathbf{e}}} \sum_{t=1}^{N}\hat{e}(t)^2/N, \tag{7a}$$

$$\text{s.t. } \hat{e}(t) - y(t) + \sum_{k=1}^{n_x} a_k\left(\hat{e}(t-k) - y(t-k)\right) + b_k u(t-k) = 0, \quad t = n_x+1,\ldots,N, \tag{7b}$$

Alternatively, one can first solve the N equality constraints in the constrained problem (6) for the N decision variables $\hat{\mathbf{e}} = \mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)^{-1}\mathbf{b}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)$ and then formulate an equivalent unconstrained optimization problem as follows:

$$\min_{\boldsymbol{\theta},\mathbf{y}_0} \mathbf{b}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)^{\mathrm{T}}\mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)^{-\mathrm{T}}\mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)^{-1}\mathbf{b}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)/N. \tag{8}$$

Although $\mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)$ is a lower triangular matrix with $\det(\mathbf{A}_{\hat{\mathbf{e}}}(\boldsymbol{\theta},\mathbf{y}_0)) = 1$, which simplifies its inversion, the cost function of this unconstrained optimization problem is typically a polynomial of high degree (up to $2N$) in $\boldsymbol{\theta}$.

Unfortunately, both (7) and (8) are nonconvex problems, which means that local optimization algorithms are prone to attain local optima and cannot guarantee global optimality of the computed solutions. Hence, it would be useful to find a method that is able to converge to the global optimum and certify this convergence. The concept of sum-of-squares polynomials that is used in this paper for global optimization is introduced in the next subsection.

B. Sum-of-squares polynomials for global optimization

This subsection recalls the concept of sum-of-squares polynomials and shows its application to global optimization.

A polynomial $p(\mathbf{x})$ of degree $2d$ in the n variables $\mathbf{x} := (x_1,\ldots,x_n)$ is a sum-of-squares (SOS) polynomial if it can be written as a sum of squares of polynomials of degree up to d in $\mathbf{x}$. The concept of SOS polynomials is useful for optimization because of the following result: $p(\mathbf{x})$ is an SOS polynomial if and only if there exists a positive semidefinite matrix $\mathbf{Q}$ such that $p(\mathbf{x}) = \mathbf{v}_d(\mathbf{x})^{\mathrm{T}}\mathbf{Q}\mathbf{v}_d(\mathbf{x}) = \mathrm{tr}\left(\mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^{\mathrm{T}}\mathbf{Q}\right)$, where $\mathbf{v}_d(\mathbf{x})$ is the $s(n,d)$-dimensional vector of monomials of degree up to d in the n variables $\mathbf{x}$, with $s(n,d) := \binom{n+d}{n}$ [15]. Hence, constraining $p(\mathbf{x})$ to the set of SOS polynomials amounts to satisfying the linear matrix inequality (LMI) $\mathbf{Q} \succeq \mathbf{0}_{s(n,d)\times s(n,d)}$, which can be done via a convex semidefinite program (SDP) [26].

An SOS polynomial $p(\mathbf{x})$ is obviously a nonnegative polynomial, that is, $p(\mathbf{x}) \geq 0\ \forall\mathbf{x}$. However, it is not generally true that a nonnegative polynomial is an SOS polynomial [27]. On the other hand, if $f(\mathbf{x})$ is a strictly positive polynomial on a compact basic semi-algebraic set $\mathbb{K}$ specified by some polynomials $g_j(\mathbf{x})$, that is, if $f(\mathbf{x}) > 0\ \forall\mathbf{x} \in \mathbb{K} = \left\{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,m\right\}$, then $f(\mathbf{x})$ can be represented as a combination of SOS polynomials provided that $\mathbb{K}$ satisfies some technical assumptions. This important result is known as Putinar's Positivstellensatz and is summarized in the following theorem [28].

Theorem 1: Consider the basic semi-algebraic set $\mathbb{K} := \{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,m\}$ and assume there exists some $q \in \{1,\ldots,m\}$ such that the set $\{\mathbf{x} : g_q(\mathbf{x}) \geq 0\}$ is compact.

If $f(\mathbf{x})$ is strictly positive $\forall \mathbf{x} \in \mathbb{K}$, then

$$f(\mathbf{x}) = p_0(\mathbf{x}) + \sum_{j=1}^{m} g_j(\mathbf{x}) p_j(\mathbf{x}) \tag{9}$$

for some SOS polynomials $p_0(\mathbf{x})$ and $p_1(\mathbf{x}),\ldots,p_m(\mathbf{x})$. ■

This representation is of interest because it can be used to relax the difficult problem of verification of positivity of $f(\mathbf{x})$ $\forall \mathbf{x} \in \mathbb{K}$ as a hierarchy of LMI feasibility problems of increasing relaxation order d [29]. To introduce the relaxations, note that the monomials $\mathbf{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ of degree up to $2d$ in the variables $\mathbf{x}$ involve powers $\boldsymbol{\alpha} := (\alpha_1,\ldots,\alpha_n)$ in the set defined as $\mathscr{X}_d := \{(\alpha_1,\ldots,\alpha_n) \in \mathbb{N}_0^n : 0 \leq \alpha_1 + \ldots + \alpha_n \leq 2d\}$. Then, we denote the coefficients of the polynomials $f(\mathbf{x})$ of degree $2v_0$ or $2v_0 - 1$ and $g_j(\mathbf{x})$ of degree $2v_j$ or $2v_j - 1$ as $f_{\boldsymbol{\alpha}}$ and $g_{j,\boldsymbol{\alpha}}$, respectively, such that $f(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathscr{X}_d} f_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ and $g_j(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathscr{X}_{v_j}} g_{j,\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$, for $j = 1,\ldots,m$, where the relaxation order $d \geq v := \max_{j=0,1,\ldots,m} v_j$. Furthermore, the matrices $\mathbf{R}_{v,\boldsymbol{\alpha}}$ are defined such that $\sum_{\boldsymbol{\alpha} \in \mathscr{X}_{d-v}} \mathbf{R}_{v,\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} = \mathbf{v}_{d-v}(\mathbf{x}) \mathbf{v}_{d-v}(\mathbf{x})^{\mathrm{T}}$, for $v = 0,\ldots,d$.

If $f(\mathbf{x})$ is strictly positive $\forall \mathbf{x} \in \mathbb{K}$, then there exists a positive integer d such that $\forall \boldsymbol{\alpha} \in \mathscr{X}_d$

$$f_{\boldsymbol{\alpha}} = \mathrm{tr}\left(\mathbf{R}_{0,\boldsymbol{\alpha}} \mathbf{Q}_0\right) + \sum_{j=1}^{m} \sum_{\substack{\boldsymbol{\beta} \in \mathscr{X}_{d-v_j} \\ \boldsymbol{\alpha}-\boldsymbol{\beta} \in \mathscr{X}_{v_j}}} g_{j,\boldsymbol{\alpha}-\boldsymbol{\beta}} \mathrm{tr}\left(\mathbf{R}_{v_j,\boldsymbol{\beta}} \mathbf{Q}_j\right) \tag{10a}$$

and

$$\mathbf{Q}_0 \succeq \mathbf{0}_{s(n,d) \times s(n,d)}, \tag{10b}$$

$$\mathbf{Q}_j \succeq \mathbf{0}_{s(n,d-v_j) \times s(n,d-v_j)},\ j = 1,\ldots,m. \tag{10c}$$

This result is very useful for global optimization. To see this, consider the problem of computing J^*, an accurate approximation of the global minimum of $J(\mathbf{x})$ subject to the constraints $g_j(\mathbf{x}) \geq 0$, for $j = 1,\ldots,m$. Equivalently, this problem can be formulated as that of computing the maximum value τ such that $f(\mathbf{x}) = J(\mathbf{x}) - \tau$ is strictly positive $\forall \mathbf{x} \in \mathbb{K} = \{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,m\}$. Using (10), such a problem can be formulated as the SDP

$$\min_{\tau, \mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_m} -\tau, \qquad \text{s.t. (10).} \tag{11}$$

Although there exist theoretical bounds for the relaxation order d that ensures the representation of a strictly positive polynomial as a combination of SOS polynomials, these bounds are unfortunately not practically useful [30]. Hence, the SDP (11) contains one LMI of size $\binom{n+d}{n}$ with $\binom{n+2d}{n}$ decision variables, where d may be very large in theory. On the other hand, in many practical situations, the relaxation order d that provides a representation in terms of SOS polynomials is not much larger than v. Furthermore, if this representation exists for some order d, a certificate can be obtained upon convergence of the SDP. The result about the representation for the order d can be stated as follows [31]:

Theorem 2: Denote the optimal values of the dual variables for the constraints (10a) as $\mu_{\boldsymbol{\alpha}}^*$ $\forall \boldsymbol{\alpha} \in \mathscr{X}_d$. If $\exists G$: $G = \mathrm{rank}\left(\sum_{\boldsymbol{\alpha} \in \mathscr{X}_d} \mathbf{R}_{0,\boldsymbol{\alpha}} \mu_{\boldsymbol{\alpha}}^*\right) = \mathrm{rank}\left(\sum_{\boldsymbol{\alpha} \in \mathscr{X}_{d-1}} \mathbf{R}_{1,\boldsymbol{\alpha}} \mu_{\boldsymbol{\alpha}}^*\right)$, then $f(\mathbf{x}) = J(\mathbf{x}) - J^*$ can be represented as in (9) with $p_0(\mathbf{x})$ of degree $2d$ and $p_j(\mathbf{x})$ of degree $2(d - v_j)$, for $j = 1,\ldots,m$. In addition, the global minimum $J^* = \tau^*$ and G global minimizers $\mathbf{x}^*$ can be computed using the fact that $\mathbf{v}_d(\mathbf{x}^*)$ lie both in the null space of $\mathbf{Q}_0^*$ and in the row space of $\mathbf{L}_0^*$, where $\mathbf{L}_0$ is the dual variable of the LMI (10b). ■

Unfortunately, for many optimization problems, the number of variables n or the maximum degree v of the polynomials in the problem may be rather large. As a result, even in the best case $d = v$, the largest LMI of size $\binom{n+d}{n}$ with $\binom{n+2d}{n}$ decision variables may be too large for computational implementation. For a problem with $n = 261$ and $d = 2$ (see an example in Section V), the LMI would be of size 34453 with 200860990 decision variables. Fortunately, a sparse version of the representation in (10) can be obtained if the original problem satisfies certain conditions. More specifically, this sparse representation takes advantage of the fact that each polynomial $g_j(\mathbf{x})$ may involve only a few variables, and $f(\mathbf{x})$ may be written as a sum of polynomials that also involve only a few variables [32]. For this, we define the p index subsets I_k with the corresponding $n_k := |I_k|$ variables $\mathbf{x}(I_k) = \{x_i : i \in I_k\}$, for $k = 1,\ldots,p$, such that $\cup_{k=1}^{p} I_k = \{1,\ldots,n\}$. This important result about sparse representation is summarized in the following theorem [33].

Theorem 3: Consider the basic semi-algebraic set $\mathbb{K} := \{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,m\}$ and assume that the index subsets $I_1,\ldots,I_p$ satisfy the following conditions:

1) The polynomial $f(\mathbf{x})$ can be written as a sum of p polynomials that involve only the variables $\mathbf{x}(I_1),\ldots,\mathbf{x}(I_p)$, that is, $f(\mathbf{x}) = \sum_{k=1}^{p} f_k(\mathbf{x}(I_k))$.
2) The running intersection property holds, that is, for all $k = 1,\ldots,p-1$, $I_{k+1} \cap \left(\cup_{j=1}^{k} I_j\right) \subseteq I_s$ for some $s \leq k$.
3) For all $j = 1,\ldots,m$, there exists some $K_j \in \{1,\ldots,p\}$ that indicates that $g_j(\mathbf{x})$ involves only the variables $\mathbf{x}(I_{K_j})$, that is, $g_j(\mathbf{x}) = g_{K_j,j}(\mathbf{x}(I_{K_j}))$.
4) For all $k = 1,\ldots,p$, there exists some $q_k \in \{1,\ldots,m\}$ such that the set $\{\mathbf{x}(I_k) : g_{k,q_k}(\mathbf{x}(I_k)) \geq 0\}$ is compact.

If $f(\mathbf{x})$ is strictly positive $\forall \mathbf{x} \in \mathbb{K}$, then

$$f(\mathbf{x}) = \sum_{k=1}^{p} p_{0,k}(\mathbf{x}(I_k)) + \sum_{j=1}^{m} g_j(\mathbf{x}) p_j(\mathbf{x}(I_{K_j})) \tag{12}$$

for some SOS polynomials $p_{0,1}(\mathbf{x}(I_1)),\ldots,p_{0,p}(\mathbf{x}(I_p))$ and $p_1(\mathbf{x}(I_{K_1})),\ldots,p_m(\mathbf{x}(I_{K_m}))$. ■

This representation may be used, in the same spirit as (9), to relax the verification of positivity of $f(\mathbf{x})$ $\forall \mathbf{x} \in \mathbb{K}$ as a hierarchy of sparse LMI feasibility problems [33]. To introduce the sparse relaxations, note that the monomials $\mathbf{x}^{\boldsymbol{\alpha}}$ of degree up to $2d$ in the variables $\mathbf{x}(I_k)$ involve powers $\boldsymbol{\alpha}$ in the set defined as $\bar{\mathscr{X}}_{d,k} := \mathscr{X}_d \cap \{(\alpha_1,\ldots,\alpha_n) \in \mathbb{N}_0^n : \alpha_i \neq 0 \Rightarrow i \in I_k\}$, for $k = 1,\ldots,p$. Then, we define the set $\bar{\mathscr{X}}_d := \cup_{k=1}^{p} \bar{\mathscr{X}}_{d,k}$ and use again $f_{\boldsymbol{\alpha}}$ and $g_{j,\boldsymbol{\alpha}}$ to denote the coefficients of the polynomials $f(\mathbf{x})$ and $g_j(\mathbf{x})$, with $\mathscr{X}_d$ in their previous definition being replaced by $\bar{\mathscr{X}}_d$. Furthermore, the

matrices $\mathbf{R}_{v,k,\boldsymbol{\alpha}}$ are defined such that $\sum_{\boldsymbol{\alpha}\in\bar{\mathscr{X}}_{d-v}} \mathbf{R}_{v,k,\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}} = \mathbf{v}_{d-v}(\mathbf{x}(I_k))\mathbf{v}_{d-v}(\mathbf{x}(I_k))^{\mathrm{T}}$, for $v = 0,\ldots,d$ and $k = 1,\ldots,p$.

If $f(\mathbf{x})$ is strictly positive $\forall \mathbf{x} \in \mathbb{K}$, then there exists a positive integer d such that $\forall \boldsymbol{\alpha} \in \bar{\mathscr{X}}_d$

$$f_{\boldsymbol{\alpha}} = \sum_{k=1}^{p} \operatorname{tr}\left(\mathbf{R}_{0,k,\boldsymbol{\alpha}}\mathbf{Q}_{0,k}\right) + \sum_{j=1}^{m} \sum_{\substack{\boldsymbol{\beta}\in\bar{\mathscr{X}}_{d-v_j} \\ \boldsymbol{\alpha}-\boldsymbol{\beta}\in\bar{\mathscr{X}}_{v_j}}} g_{j,\boldsymbol{\alpha}-\boldsymbol{\beta}} \operatorname{tr}\left(\mathbf{R}_{v_j,K_j,\boldsymbol{\beta}}\mathbf{Q}_j\right) \tag{13a}$$

and

$$\mathbf{Q}_{0,k} \succeq \mathbf{0}_{s(n_k,d)\times s(n_k,d)},\ k = 1,\ldots,p, \tag{13b}$$

$$\mathbf{Q}_j \succeq \mathbf{0}_{s(n_{K_j},d-v_j)\times s(n_{K_j},d-v_j)},\ j = 1,\ldots,m. \tag{13c}$$

Hence, when the conditions of Theorem 3 are satisfied, (11) is equivalent to

$$\min_{\tau,\mathbf{Q}_{0,1},\ldots,\mathbf{Q}_{0,p},\mathbf{Q}_1,\ldots,\mathbf{Q}_m} -\tau, \qquad \text{s.t. (13).} \tag{14}$$

To appreciate the advantage of the sparse representation (13) over (10), suppose that n_k is the same for all $k = 1,\ldots,p$, with $n_k = \lfloor \frac{n}{p} \rfloor + n \bmod p$. Then, the SDP (14) contains p LMIs of size $\binom{n_k+d}{n_k}$ with $\binom{n_k+2d}{n_k}$ decision variables instead of one LMI of size $\binom{n+d}{n}$ with $\binom{n+2d}{n}$ decision variables in (11). For the aforementioned problem with $n = 261$ and $d = 2$, if $p = 252$ (which is also the case in Section V), then the SDP contains 252 LMIs of size 66 with 1001 decision variables instead of one LMI of size 34453 with 200860990 decision variables. Hence, if n_k and the maximum degree v of the polynomials in the problem are relatively small, the SDP can be solved efficiently since the relaxation order d that provides a sparse representation in terms of SOS polynomials is usually not much larger than v. If this representation exists for some order d, a certificate can again be obtained upon convergence of the SDP. The result about the sparse representation for the order d can be stated as follows [31]:

Theorem 4: Denote the optimal values of the dual variables for the constraints (13a) as $\mu^*_{\boldsymbol{\alpha}}$ $\forall \boldsymbol{\alpha} \in \bar{\mathscr{X}}_d$. If $\exists G : G = \operatorname{rank}\left(\sum_{\boldsymbol{\alpha}\in\bar{\mathscr{X}}_d} \mathbf{R}_{0,k,\boldsymbol{\alpha}}\mu^*_{\boldsymbol{\alpha}}\right) = \operatorname{rank}\left(\sum_{\boldsymbol{\alpha}\in\bar{\mathscr{X}}_{d-1}} \mathbf{R}_{1,k,\boldsymbol{\alpha}}\mu^*_{\boldsymbol{\alpha}}\right)$ $\forall k = 1,\ldots,p$, then $f(\mathbf{x}) = J(\mathbf{x}) - J^*$ can be represented as in (12) with $p_{0,k}(\mathbf{x}(I_k))$ of degree $2d$, for $k = 1,\ldots,p$, and $p_j(\mathbf{x}(I_{K_j}))$ of degree $2(d-v_j)$, for $j = 1,\ldots,m$. In addition, the global minimum $J^* = \tau^*$ and G global minimizers $\mathbf{x}^*$ can be computed using the fact that $\mathbf{v}_d(\mathbf{x}(I_k)^*)$ lie both in the null space of $\mathbf{Q}^*_{0,k}$ and in the row space of $\mathbf{L}^*_{0,k}$, where $\mathbf{L}_{0,k}$ is the dual variable of the LMI (13b), $\forall k = 1,\ldots,p$. ■

III. GLOBAL SOLUTIONS TO ML ESTIMATION PROBLEMS FOR OE LINEAR MODELS

This section shows how to apply the concept of SOS polynomials presented in Section II-B to obtain global solutions to the ML estimation problems described in Section II-A.

In Section II-A, it has been shown that the ML parameter estimation problem for the model in (3) can be formulated as the constrained problem (7) or the unconstrained problem (8). In terms of the notation in Section II-B, the constrained problem involves $n = N + 2n_x$ decision variables, and each polynomial in the problem, both in the cost function and the constraints, is at most of degree 2, which means that $v = 1$, whereas the unconstrained problem involves $n = 3n_x$ decision variables, and the only polynomial in the problem, the cost function, is of degree $2N$, which means that $v = N$. Since each relaxation order d in the hierarchy of nonsparse semidefinite relaxations requires solving one LMI of size $\binom{n+d}{n}$, with $d \geq v$, both problem formulations become intractable for a large sample size N.

However, one can note that, in the constrained problem (7), each equality constraint corresponds to a quadratic polynomial that involves only the $2n_x$ variables $\boldsymbol{\theta}$ and n_x+1 variables from $\hat{\mathbf{e}}$, and the cost function can be written as a sum of quadratic polynomials that involve only a few variables from $\hat{\mathbf{e}}$. This allows the use of a hierarchy of sparse semidefinite relaxations if each equality constraint is transformed into a pair of inequality constraints to obtain a basic semi-algebraic set.

Hence, we introduce the following definitions:

$$f(\mathbf{x}) := J(\mathbf{x}) - \tau, \tag{15a}$$

$$g_j(\mathbf{x}) := \begin{cases} -h_{j+n_x}(\mathbf{x}),\ j = 1,\ldots,N-n_x, \\ h_{j-N+2n_x}(\mathbf{x}),\ j = N-n_x+1,\ldots,2N-2n_x, \end{cases} \tag{15b}$$

with $\mathbf{x} := (\boldsymbol{\theta},\hat{\mathbf{e}}) = (a_1,\ldots,a_{n_x},b_1,\ldots,b_{n_x},\hat{e}(1),\ldots,\hat{e}(N))$ and

$$J(\mathbf{x}) = \sum_{t=1}^{N} \hat{e}(t)^2/N, \tag{15c}$$

$$h_t(\mathbf{x}) = \hat{e}(t) - y(t) + \sum_{k=1}^{n_x} a_k\left(\hat{e}(t-k) - y(t-k)\right) + b_k u(t-k), \quad t = n_x+1,\ldots,N. \tag{15d}$$

Then, the problem (7) is equivalent to computing the maximum τ such that $f(\mathbf{x})$ is strictly positive $\forall \mathbf{x} \in \mathbb{K} = \left\{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,2N-2n_x\right\}$. The previous definitions seem to suggest the choice of $n_k = 3n_x+1$ variables $\mathbf{x}(I_k) = (a_1,\ldots,a_{n_x},b_1,\ldots,b_{n_x},\hat{e}(k),\ldots,\hat{e}(k+n_x))$ and corresponding index subsets $I_k = \{1,\ldots,2n_x,k+2n_x,\ldots,k+3n_x\}$, for $k = 1,\ldots,p$, with $p := N - n_x$. We now show that the conditions in Theorem 3 are satisfied for these index subsets.

Condition 1 is satisfied by using $f_1(\mathbf{x}(I_1)) = \hat{e}(1)^2/N + \ldots + \hat{e}(n_x+1)^2/N - \tau$ and $f_k(\mathbf{x}(I_k)) = \hat{e}(n_x+k)^2/N$ for $k = 2,\ldots,p$. The running intersection property in Condition 2 is also satisfied since, for all $k = 1,\ldots,p-1$, $I_{k+1} \cap \left(\cup_{j=1}^{k} I_j\right) = \{1,\ldots,2n_x,k+1+2n_x,\ldots,k+3n_x\} \subseteq I_k$. Condition 3 is satisfied by construction since, for all $j = 1,\ldots,N-n_x$, $g_j(\mathbf{x})$ involves only the variables $\mathbf{x}(I_{K_j})$, with $K_j = j$, and, for all $j = N-n_x+1,\ldots,2N-2n_x$, $g_j(\mathbf{x})$ involves only the variables $\mathbf{x}(I_{K_j})$, with $K_j = j-N+n_x$.

Unfortunately, Condition 4 is not satisfied initially, but it is possible to add additional constraints to ensure that it is satisfied. Hence, we redefine $\mathbb{K} = \left\{\mathbf{x} : g_j(\mathbf{x}) \geq 0, \forall j = 1,\ldots,m\right\}$, with $m := 3N - 3n_x$, by adding the quadratic polynomials

$$g_j(\mathbf{x}) := -\bar{h}_{j-2N+3n_x}(\mathbf{x}),\ j = 2N-2n_x+1,\ldots,m, \tag{15e}$$

with

$$\bar{h}_t(\mathbf{x}) = -r^2 + \hat{e}(t)^2 + \sum_{k=1}^{n_x} \left(a_k^2 + b_k^2 + \hat{e}(t-k)^2\right), \quad t = n_x+1,\ldots,N, \qquad (15f)$$

where r is some finite constant. It is important to observe that, if r is chosen large enough to ensure that the minimizers $\mathbf{x}^*$ of problem (7) are such that $||\mathbf{x}(I_k)^*|| \leq r$, for $k = 1,\ldots,p$, then the new constraints are redundant because adding them does not change the minimizers. Moreover, the polynomials (15f) are chosen to be quadratic since the polynomials with compact superlevel sets are at least of degree 2 and the polynomials that specify the other constraints are also of degree $2v_j = 2$. Then, Condition 3 is still satisfied with the new constraints since, for all $j = 2N-2n_x+1,\ldots,3N-3n_x$, $g_j(\mathbf{x})$ involves only the variables $\mathbf{x}(I_{K_j})$, with $K_j = j-2N+2n_x$. In addition, now Condition 4 is also satisfied since, for all $k = 1,\ldots,p$, the superlevel set $\left\{\mathbf{x}(I_k) : g_{k,q_k}(\mathbf{x}(I_k)) \geq 0\right\} = \{\mathbf{x}(I_k) : ||\mathbf{x}(I_k)|| \leq r\}$ is compact for $q_k = k+2N-2n_x$.

Some comments about the boundedness of $||\mathbf{x}(I_k)^*||$, for $k = 1,\ldots,p$, are necessary at this point. Since $\mathbf{x}(I_k) = (a_1,\ldots,a_{n_x},b_1,\ldots,b_{n_x},\hat{e}(k),\ldots,\hat{e}(k+n_x))$, this boundedness implies that the parameters $\boldsymbol{\theta}$ and n_x+1 prediction errors from $\hat{\mathbf{e}}$ are bounded. At least in the case of bounded-input bounded-output (BIBO) stable systems, it seems reasonable to assume that the parameters $\boldsymbol{\theta}$ are bounded. Regarding the prediction errors, they are expected to have the same magnitude as the output errors, which are assumed to be realizations of a normally distributed random variable with zero mean and variance σ^2 in this paper. Although in theory the support of this random variable is unbounded, in practice it can be bounded with a very high confidence level. To be more precise, one can observe that (15f) includes the sum of squares of n_x+1 predicted errors $\hat{w}(t)^2 := \sum_{k=0}^{n_x} \hat{e}(t-k)^2$ that corresponds to the sum of squares of n_x+1 output errors $w(t)^2 := \sum_{k=0}^{n_x} e(t-k)^2$, for $t = n_x+1,\ldots,N$. Then, $\frac{w(t)^2}{\sigma^2}$ is the realization of a random variable W that follows a chi-squared distribution with n_x+1 degrees of freedom, and the probability that $w(t)^2 \leq r_w^2$ is $F_W(\frac{r_w^2}{\sigma^2})$, where F_W denotes the cumulative density function of the random variable W. Hence, we propose the bound $r^2 = r_{||\boldsymbol{\theta}||}^2 + r_{\hat{w}}^2$ for (15f), where $r_{||\boldsymbol{\theta}||}^2$ is an upper bound on $||\boldsymbol{\theta}||^2$ and $r_{\hat{w}}^2$ is an upper bound on $\hat{w}(t)^2$ that is chosen according to the desired robustness.

Since all the conditions in Theorem 3 are satisfied, the problem that consists in computing the global minimum of $J(\mathbf{x})$ subject to $g_j(\mathbf{x}) \geq 0$, for $j = 1,\ldots,m$, can be formulated as the SDP (14) as described in Section II-B. A certificate of the representation in terms of SOS polynomials for some order d can be obtained upon convergence of the SDP as shown in Theorem 4, which is a certificate of global optimality of the solution $\mathbf{x}^* := (\boldsymbol{\theta}^*, \hat{\mathbf{e}}^*)$ and the MSE τ^*.

Suppose that a global optimum is computed and certified for the relaxation order $d = 2$ (in fact, this is always the case in the example of Section V). This implies that the SDP (14) has been solved for $d = 2$, which is an SDP with $p\binom{n_k+2d}{n_k} - (p-1)\binom{|I_k \cap I_{k+1}|+2d}{|I_k \cap I_{k+1}|} = (4N-n_x+1)\frac{(3n_x+2)(3n_x+3)(3n_x+4)}{24}$ equality constraints, $p = N-n_x$ LMIs of size $\binom{n_k+d}{n_k} = \frac{(3n_x+2)(3n_x+3)}{2}$, and $m = 3N-3n_x$ LMIs of size $\binom{n_k+d-v_j}{n_k} = 3n_x+2$. Note that, thanks to the sparse representation, the input size of this SDP is linear in the sample size N, which would not be possible with the nonsparse representation. Since the complexity of SDPs is polynomial in their input size, it means that it has been possible to compute and certify a global solution $\mathbf{x}^*$ in polynomial time.

IV. EXTENSION TO OTHER MODELS

Besides the discrete-time OE LTI SISO models that were considered previously, the idea of using the concept of SOS polynomials and sparse semidefinite relaxations may be beneficial for ML parameter estimation of other model structures. For example, the extension of this idea to discrete-time OE LTI multiple-input single-output (MISO) models seems to be straightforward, although its analysis is not detailed here. This idea could also be extended to models that are nonlinear but share the same rational structure of linear models described by (1). For example, this class of models is used to represent the Monod-type kinetics that is frequently used in models of biological systems to account for activating or inhibitory effects of certain chemical species [34]. However, this extension is not detailed here.

V. SIMULATION EXAMPLE

In this section, we consider a 2-mass-spring-damper system with a single unconstrained rigid-body degree of freedom adapted from [35]. The implementation was performed on MATLAB R2018a running on an Intel Core i7 1.9 GHz processor. MOSEK 8.1 was used as SDP solver.

In this example, the input is the actuation force on the first mass and the output is the velocity of the first mass. The continuous-time transfer function of the system is

$$P_c(s) = c\frac{m_2 s^2 + ds + k}{m_1 m_2 s^3 + (m_1+m_2)ds^2 + (m_1+m_2)ks}, \qquad (16)$$

with $[c,k,d,m_1,m_2] = \left[56, 1.8\times10^4, 2.0, 0.1, 1.9\right]$. By using zero-order hold for discretization and a sampling period $h = 0.004$ s, this continuous-time transfer function is converted into the following discrete-time transfer function:

$$P_d(z) = \frac{1.2673z^{-1} - 2.2742z^{-2} + 1.2582z^{-3}}{1 - 0.6759z^{-1} + 0.5951z^{-2} - 0.9192z^{-3}}. \qquad (17)$$

For system identification, a pseudo-random binary signal (PRBS) of size N in a range between -1 and 1 is applied as the input of this system of order $n_x = 3$. The output is corrupted by additive i.i.d. Gaussian noise with the standard deviation $\sigma = 0.1$. Then, the input and output data are used to formulate the SDP (14), where $f(\mathbf{x})$ and $g_j(\mathbf{x})$, for $j = 1,\ldots,m$, are given in (15). For each $N \in \{31,63,127,255\}$, 100 repetitions of this procedure are performed, with different realizations of the noise for each repetition.

In all the repetitions, it is possible to extract the unique solution $\boldsymbol{\theta}^* = (a_1^*, a_2^*, a_3^*, b_1^*, b_2^*, b_3^*)$ from the solution to the SDP for the relaxation order $d = 2$ and certify the global optimality of the solution $\boldsymbol{\theta}^*$ that corresponds to the MSE τ^*.

TABLE I

EXECUTION TIME IN SECONDS, MSE τ^*, AND ESTIMATES a_1^*, a_2^*, a_3^*, b_1^*, b_2^*, b_3^* (MEAN $\pm$ STANDARD DEVIATION FOR 100 REPETITIONS) OF GLOBAL ML PARAMETER ESTIMATION FOR DIFFERENT SAMPLE SIZES N.

N	Time (s)	τ^*	a_1^*	a_2^*	a_3^*	b_1^*	b_2^*	b_3^*
31	19.2 ± 2.5	$0.72 \pm 0.22 \times 10^{-2}$	-0.6753 ± 0.0073	0.5950 ± 0.0030	-0.9186 ± 0.0067	1.2667 ± 0.0193	-2.2722 ± 0.0184	1.2586 ± 0.0221
63	40.3 ± 4.3	$0.87 \pm 0.18 \times 10^{-2}$	-0.6759 ± 0.0010	0.5953 ± 0.0008	-0.9194 ± 0.0009	1.2657 ± 0.0106	-2.2719 ± 0.0135	1.2594 ± 0.0109
127	82.7 ± 8.5	$0.94 \pm 0.13 \times 10^{-2}$	-0.6759 ± 0.0005	0.5951 ± 0.0005	-0.9192 ± 0.0004	1.2669 ± 0.0059	-2.2734 ± 0.0064	1.2582 ± 0.0062
255	194.4 ± 24.0	$0.97 \pm 0.09 \times 10^{-2}$	-0.6759 ± 0.0003	0.5951 ± 0.0003	-0.9192 ± 0.0002	1.2682 ± 0.0034	-2.2744 ± 0.0045	1.2580 ± 0.0042

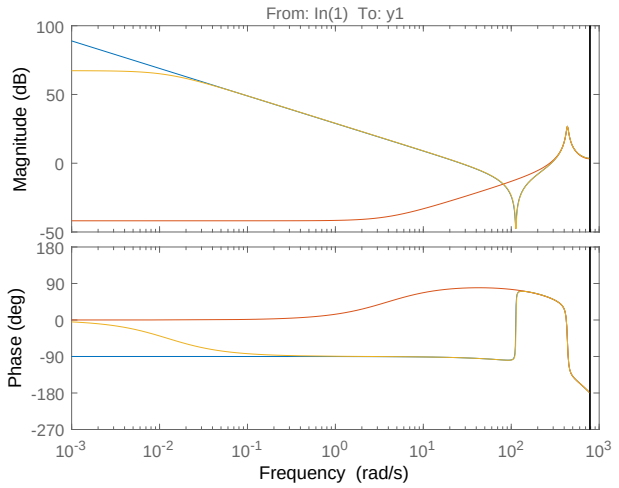

Fig. 1. Bode diagram for one of the repetitions with the sample size $N = 255$. The blue line corresponds to the true model, the yellow line corresponds to the model estimated via the proposed approach, and the red line corresponds to the model estimated via `oe`.

Note that a different solution with a larger MSE is provided in some cases by the `oe` function from the MATLAB System Identification Toolbox when it is initialized with the parameters that result from the use of subspace identification via the `n4sid` function from the same toolbox [36]. More precisely, this occurs in the case of 19 repetitions for $N = 31$, 5 repetitions for $N = 63$, 3 repetitions for $N = 127$, and 2 repetitions for $N = 255$. This indicates that `oe` is reporting local minima in these cases. Fig. 1 compares the Bode diagram for the true model and the models estimated via the proposed approach and via `oe` for one of these cases. For example, the model estimated via the proposed approach recovers the antiresonance peak, while the model estimated via `oe` is not able to do so. Hence, the advantage of the proposed approach is that it avoids local minima and certifies the global optimality of the computed solution.

Table I reports the execution time of the whole procedure for global optimization, the MSE, and the parameter estimates for the different sample sizes N. The duration of the pre-processing steps (formulation of the SDP) and the post-processing steps (extraction and certification of the global solution) is much smaller than the execution time of the SDP solver. It is possible to observe that the execution time seems to be approximately a linear function of N, the MSE converges to its expected value of $\sigma^2 = 10^{-2}$ for large N, and the parameter estimates converge to the true parameters.

VI. CONCLUSIONS

This paper has shown that the concept of SOS polynomials and the resulting hierarchy of sparse semidefinite relaxations can be used for tractable computation of ML parameter estimates *with posterior certification of global optimality*. This computation and certification has been shown in detail for the case of discrete-time OE LTI SISO models, but it has also been noted that the methodology can be extended to a much broader family of models described by rational structures, in particular some nonlinear models. The use of sparse semidefinite relaxations for ML parameter estimation takes advantage of the sparse structure of the optimization problem that results from the ML method for certain rational model structures. These features have been illustrated by a simulation example of a third-order LTI SISO system.

Future work may focus on improving the efficiency of the method, for example by finding ways to further increase the sparsity of the estimation problem, and extending this method to the ML parameter estimation of other linear and nonlinear models. For example, it would be interesting to investigate the potential of the method for stochastic Wiener-Hammerstein models [37] and nonlinear model structures of biological systems described by Monod terms [34].

REFERENCES

[1] Y. Bard, *Nonlinear Parameter Estimation*. New York, NY: Academic Press, 1974.

[2] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, 1977.

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

[4] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*, 2nd ed. Wiley, 2012.

[5] T. Makelainen, K. Schmidt, and G. P. H. Styan, "On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples," *Ann. Statist.*, vol. 9, pp. 758–767, 1981.

[6] A. Wills, C. Yu, L. Ljung, and M. Verhaegen, "Affinely parametrized state-space models: Ways to maximize the likelihood function," in *Proc. 18th IFAC Symposium on System Identification*, Stockholm, Sweden, Jul. 2018, pp. 718–723.

[7] C. Yu, L. Ljung, and M. Verhaegen, "Identification of structured state-space models," *Automatica*, vol. 90, pp. 54 – 61, 2018.

[8] N. Everitt, M. Galrinho, and H. Hjalmarsson, "Open-loop asymptotically efficient model reduction with the Steiglitz-McBride method," *Automatica*, vol. 89, pp. 221 – 234, 2018.

[9] M. Galrinho, C. R. Rojas, and H. Hjalmarsson, "Parametric identification using weighted null-space fitting," *IEEE Trans. Automat. Contr.*, 2019, to be published.

[10] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000.

[11] J. Clausen, "Branch and bound algorithms - Principles and examples," University of Copenhagen, Tech. Rep., 1999.

[12] K. Batselier and B. De Moor, "Maximum likelihood and polynomial system solving," in *Proc. 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Hong Kong, Dec. 2010, pp. 819–820.

[13] K. Batselier, P. Dreesen, and B. De Moor, "Prediction error method identification is an eigenvalue problem," in *Proc. 16th IFAC Symposium on System Identification*, Brussels, Belgium, Jul. 2012, pp. 221–226.

[14] M. Laurent, "Sums of squares, moment matrices and optimization over polynomials," in *Emerging Applications of Algebraic Geometry*, M. Putinar and S. Sullivant, Eds. New York, NY: Springer, 2009, pp. 157–270.

[15] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications*. London, UK: Imperial College Press, 2010.

[16] S. Prajna, P. A. Parrilo, and A. Rantzer, "Nonlinear control synthesis by convex optimization," *IEEE Trans. Automat. Contr.*, vol. 49, no. 2, pp. 310–314, Feb. 2004.

[17] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat, "Nonlinear optimal control via occupation measures and LMI-relaxations," *SIAM J. Control Optim.*, vol. 47, no. 4, pp. 1643–1666, 2008.

[18] D. Henrion and M. Korda, "Convex computation of the region of attraction of polynomial control systems," *IEEE Trans. Automat. Contr.*, vol. 59, pp. 297–312, 2014.

[19] N. Ozay, C. Lagoa, and M. Sznaier, "Robust identification of switched affine systems via moments-based convex optimization," in *Proc. 48h IEEE Conference on Decision and Control (CDC)*, Shanghai, PR China, Dec. 2009, pp. 4686–4691.

[20] M. Ayazoglu, M. Sznaier, C. Lagoa, and O. Camps, "A moments-based approach to estimation and data interpolation for a class of Wiener systems," in *Proc. 49th IEEE Conference on Decision and Control (CDC)*, Atlanta, GA, Dec. 2010, pp. 5674–5680.

[21] J. Mårtensson and H. Hjalmarsson, "Robust input design using sum of squares constraints," in *Proc. 14th IFAC Symposium on System Identification*, Newcastle, Australia, Mar. 2006, pp. 1352–1357.

[22] P. A. Parrillo and L. Ljung, "Initialization of physical parameter estimates," in *Proc. 13th IFAC Symposium on System Identification*, Rotterdam, The Netherlands, Aug. 2003, pp. 1483–1488.

[23] V. Cerone, J. B. Lasserre, D. Piga, and D. Regruto, "A unified framework for solving a general class of conditional and robust set-membership estimation problems," *IEEE Trans. Automat. Contr.*, vol. 59, no. 11, pp. 2897–2909, Nov. 2014.

[24] K. J. Åström and T. Bohlin, "Numerical identification of linear dynamic systems from normal operating records," in *Theory of Self-Adaptive Control Systems*. Plenum Press, Jan. 1966, pp. 96–111.

[25] T. Söderström and P. Stoica, *System Identification*. Hemel Hempstead, UK: Prentice Hall International, 1989.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[27] T. S. Motzkin, "The arithmetic-geometric inequality," in *Proc. Symposium on Inequalities*, New York, NY, 1967, pp. 205–224.

[28] M. Putinar, "Positive polynomials on compact semi-algebraic sets," *Ind. Univ. Math. J.*, vol. 42, pp. 969–984, 1993.

[29] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, pp. 796–817, 2001.

[30] J. Nie and M. Schweighofer, "On the complexity of Putinar's positivstellensatz," *J. Complexity*, vol. 23, pp. 135–150, 2007.

[31] J. B. Lasserre, "A semidefinite programming approach to the generalized problem of moments," *Math. Program.*, vol. 112, pp. 65–92, 2008.

[32] M. Kojima, S. Kim, and M. Maramatsu, "Sparsity in sums of squares of polynomials," *Math. Program.*, vol. 103, pp. 45–62, 2005.

[33] J. B. Lasserre, "Convergent SDP-relaxations in polynomial optimization with sparsity," *SIAM J. Optim.*, vol. 17, pp. 822–843, 2006.

[34] G. Bastin and D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors*. Amsterdam: Elsevier, 1990.

[35] R. Voorhoeve, A. V. der Maas, and T. Oomen, "Non-parametric identification of multivariable systems: A local rational modeling approach with application to a vibration isolation benchmark," *Mech. Syst. Signal Process.*, vol. 105, pp. 129–152, 2018.

[36] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, pp. 75–93, 1994.

[37] M. R. Abdalmoaty and H. Hjalmarsson, "Application of a linear PEM estimator to a stochastic Wiener-Hammerstein benchmark problem," in *Proc. 18th IFAC Symposium on System Identification*, Stockholm, Sweden, Jul. 2018, pp. 784–789.