Identification of linear time-invariant systems from multiple experiments

I. Markovsky and R. Pintelon

Abstract

A standard assumption for consistent estimation in the errors-in-variables setting is persistency of excitation of the noise free input signal. We relax this assumption by considering data from multiple experiments. Consistency is obtained asymptotically as the number of experiments tends to infinity. The main theoretical and algorithmic difficulties are related to the growing number of to-be-estimated initial conditions. The method proposed in the paper is based on analytic elimination of the initial conditions and optimization over the remaining parameters. The resulting estimator is consistent, however, achieving asymptotically efficiency remains an open problem.

Index Terms

maximum likelihood system identification, sum-of-damped exponentials modeling, consistency, structured low-rank approximation.

EDICS: SSP-IDEN, SSP-PARE, SSP-SYSM

I. Introduction

Although in static estimation problems the data is collected from repeated experiments, the default setting in system identification is data consisting of a single trajectory. The rationale for this is that more data can be obtained by increasing the measurement time instead of the number of experiments. Then, as the measurement time tends to infinity, under suitable assumptions it is possible to estimate consistently the model parameters.

In some cases, however, increasing the measurement time is either not possible or does not allow consistent estimation. When the input is given and fixed (can not be chosen) and the system is unstable, the output can be exponentially growing. This essentially restricts the measurement time. In a sense, the opposite case is when the system is stable and autonomous. Then, the true response decays exponentially, so that with a fixed noise power, the signal-to-noise ratio is diminishing as the measurement time is

I. Markovsky and R. Pintelon are with the Department ELEC, Vrije Universiteit Brussel, 1050 Brussels, Belgium (e-mail: ivan.markovsky@vub.ac.be, rik.pintelon@vub.ac.be)

growing. Indeed, an assumption required for identifiability of the system is that the input is persistently exciting [1] and the autonomous system does not satisfy this assumption. We show that consistency can be achieved without a persistently exciting input by collecting data from multiple experiments.

The multiple experiments identification problem is considered in [2] where the trajectories of the different experiments are concatenated into a single long trajectory. The transition from one trajectory to another is taken into account by including pulses in the input at the transition times. Identification from multiple experiments is implemented in the System Identification Toolbox of Matlab. The prediction error methods [3], [4] perform the optimization over the model parameters and the set of the initial conditions for all experiments.

In [5] as well as in the System Identification Toolbox, the initial conditions are treated as part of the model parameters. Consequently the nonlinear optimization problem for the prediction error minimization is solved over an increasing number of optimization variables. As a result the computational cost grows cubically in the number of the experiments. This makes the identification from a large number of experiments unpractical.

System identification in the errors-in-variables setting from multiple experiments is considered in [6] in the setting of structured low-rank approximation [7], [8], [9]. The initial conditions as well as the estimated inputs are eliminated from the optimization problem analytically, reducing the problem to optimization over the model parameters only. In this way the number of optimization variables for the nonlinear optimization is independent of the number of experiments. It is shown in [10] that the elimination step can be done efficiently with computational cost that is linear in the number of experiments. Linear cost and readily available software [11] make the identification from a large number of experiments as cheap and easily as identification from a single long trajectory.

The approach used in [6] to eliminate analytically the initial conditions is known in the numerical linear algebra literature as the variable projections method [12] for solution of separable nonlinear least squares problems. The nonlinear optimization problem problem is solved with the Levenberg-Marquardt algorithm, using the pseudo-Jacobian [13]. In this paper we prove that the approach of [6] leads to a consistent estimation method.

The paper is organized as follows. In Section II we introduce the notation. The considered identification problem from multiple experiments is defined in Section III. Section IV presents the variable projection method. The main results of the paper—the consistency proof of the variable projections method—is given in Section V. A special case of the results—consistent estimation of autonomous systems in the output error setup—is presented in Section VI. Section VII illustrates the theoretical results on a numerical example.

II. Notation

We use the behavioral language [14], [15]. The class of linear time-invariant systems with $\mathtt{m}$ inputs is denoted by $\mathscr{L}_{\mathtt{m}}$. A system $\mathscr{B} \in \mathscr{L}_{\mathtt{m}}$ is a set of trajectories [15]. The statement $w \in \mathscr{B}$ is a short-hand notation for "w is a trajectory of $\mathscr{B}$". There is a non-unique partitioning of the variables into inputs and outputs, however, the number of inputs and the number of outputs are invariant of the system and do not depend on the choice of the input-output representation.

A discrete-time system $\mathscr{B}$ is a collection of vector time-series $w : \mathbb{Z} \to \mathbb{R}^q$, where q is the number of variables: $\mathtt{m}$ inputs and $\mathtt{p} := q - \mathtt{m}$ outputs. A discrete-time linear time-invariant system $\mathscr{B} \in \mathscr{L}_{\mathtt{m}}^q$ admits a representation (refered to as a kernel representation) by a constant coefficients difference equation

$$\mathscr{B}(R) := \{\, w \mid R_0 w + R_1 \sigma w + \cdots + R_\ell \sigma^\ell w = 0 \,\}, \tag{KER}$$

where σ is the shift operator

$$(\sigma w)(t) = w(t+1).$$

It can be written more compactly as the kernel

$$\ker\big(R(\sigma)\big) := \{\, w \mid R(\sigma) w = 0 \,\}$$

of the operator $R(\sigma)$, where

$$R(z) := R_0 z + R_1 z + \cdots + R_\ell z^\ell$$

is a polynomial matrix. The minimal natural number ℓ, for which there exists an ℓth order difference equation representation for $\mathscr{B}$ is an invariant of the system, called the *lag*. Let $w_{\mathrm{p}} \wedge w_{\mathrm{f}}$ be the concatenation of the trajectories w_{p} and w_{f}. The restriction of $\mathscr{B}$ to the interval $[1, T]$ is the set of all T-samples long trajectory of $\mathscr{B}$, *i.e.*,

$$\mathscr{B}_T := \{\, w = \big(w(1), \ldots, w(T)\big) \mid \text{there are } w_{\mathrm{p}} \text{ and } w_{\mathrm{f}}, \quad \text{such that } w_{\mathrm{p}} \wedge w \wedge w_{\mathrm{f}} \in \mathscr{B} \,\}.$$

For $T \geq \ell$, the dimension of $\mathscr{B}_T$ is bounded by $T\mathtt{m} + \ell(q - \mathtt{m})$. The subset of linear time-invariant systems $\mathscr{L}_{\mathtt{m}}$ with lag at most ℓ is denoted by $\mathscr{L}_{\mathtt{m},\ell}$.

III. Problem formulation

Consider N trajectories

$$\overline{w}^i = \big(\overline{w}^i(1), \ldots, \overline{w}^i(T_i)\big), \qquad i = 1, \ldots, N,$$

with possibly different lengths $T_1, \ldots, T_N$ of a system $\overline{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$. The data

$$\mathscr{D} := \{\, w^1, \ldots, w^N \,\}$$

for the identification problem considered in the paper is generated in the errors-in-variables setup:

$$w^i = \overline{w}^i + \widetilde{w}^i, \quad \text{where } \overline{w}^i \in \overline{\mathscr{B}}_{T_i},\ \overline{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$$

$$\begin{aligned} &\widetilde{w}^i \text{ is zero mean white Gaussian} \\ &\text{process with covariance } s^2 I \\ &\text{and } \widetilde{w}^i \text{ is independent of } \widetilde{w}^j \\ &\text{for all } i \neq j. \end{aligned} \tag{EIV}$$

Here $\overline{w}^i$ is the "true value" of the trajectory w^i and $\overline{\mathscr{B}}$ is refered to as the "true system". In addition, we assume that there are scalars c_1 and c_2, such that

$$0 < c_1 \leq \|\overline{w}^i\|_2^2 \leq c_2 < \infty, \quad \text{for} \quad i = 1, \dots, N. \tag{A}$$

Our aim is to estimate the true system $\overline{\mathscr{B}}$ from the data $\mathscr{D}$ and the prior knowledge that the true system belongs to the model class $\mathscr{L}_{\mathtt{m},\ell}$.

Problem 1. (Maximum likelihood identification from multiple trajectories) Given a set of trajectories $\mathscr{D}$ and a model class $\mathscr{L}_{\mathtt{m},\ell}$, specified by the natural numbers $\mathtt{m}$ and ℓ, find a maximum likelihood estimate $\widehat{\mathscr{B}}$ of the true data generating system $\overline{\mathscr{B}}$.

The log likelihood function for the data generating model (EIV) is

$$L(\widehat{\mathscr{B}}, \widehat{\mathscr{D}}) = \begin{cases} \text{const} - \dfrac{1}{2s^2} \displaystyle\sum_{i=1}^{N} \|w^i - \widehat{w}^i\|_2^2 & \text{if } \widehat{w}^i \in \widehat{\mathscr{B}}_{T_i}, \\ & \text{for } i = 1, \dots, N \\ -\infty & \text{otherwise.} \end{cases}$$

The maximum likelihood principle leads to the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{N} \sum_{i=1}^{N} \|w^i - \widehat{w}^i\|_2^2 \quad \text{over } \widehat{w}^1, \dots, \widehat{w}^N \text{ and } \widehat{\mathscr{B}} \\ &\text{subject to} \quad \widehat{w}^i \in \widehat{\mathscr{B}}_{T_i}, \text{ for } i = 1, \dots, N, \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}. \end{aligned} \tag{ML}$$

Note 2 (Parameterization of a trajectory by input and initial conditions). The condition $\widehat{w}^i \in \widehat{\mathscr{B}}_{T_i}$ implies that in an input/output partitioning of the variables $w = (u, y)$, there is an input $\widehat{u}^i$ and initial conditions

$$\widehat{w}^i_{\text{ini}} = \big(\widehat{w}^i_{\text{ini}}(-\ell+1), \dots, \widehat{w}^i_{\text{ini}}(-1), \widehat{w}^i_{\text{ini}}(0)\big),$$

such that $\widehat{y}^i$ is the response of $\widehat{\mathscr{B}}$, generated by $\widehat{u}^i$, under the initial conditions $\widehat{w}^i_{\text{ini}}$. As shown in the next section, $\widehat{w}^1, \dots, \widehat{w}^N$ (and therefore the corresponding inputs and initial conditions) can be eliminated analytically, resulting in an equivalent optimization problem over $\widehat{\mathscr{B}}$ only.

Note 3 (Parameterization of the model by a representation). A finite dimensional linear time-invariant model $\mathscr{B}$ admits many representations, *e.g.*, the kernel representation (KER). Once a representation is chosen the model can be associated with the parameters of the representation, *e.g.*, the polynomial matrix $R(z)$ in the case of the kernel representation. Consequently, the likelihood L can be written as a function of the model parameters and the maximum likelihood estimation problem (ML) becomes a parameter optimization problem.

Following the behavioral approach, we formulate the problem and state our results without involving a model representation. (The numerical solution of the problem, however, requires a choice of a particular model representation.) The decision to work in the behavioral setting posses the question of what means that a model converges to another model. This question is handled in [16] by defining the convergence

$$\widehat{\mathscr{B}} \to \bar{\mathscr{B}} \qquad \iff \qquad \angle(\bar{\mathscr{B}}, \widehat{\mathscr{B}}) \to 0$$

in the sense of the gap metric $\angle(\bar{\mathscr{B}}, \widehat{\mathscr{B}})$, see [17].

IV. Elimination of $\widehat{w}^1, \ldots, \widehat{w}^N$

In this section we consider the problem of evaluating the likelihood

$$M(\widehat{\mathscr{B}}, \mathscr{D}) = \min_{\widehat{w}^1 \in \widehat{\mathscr{B}}_{T_1}, \ldots, \widehat{w}^N \in \widehat{\mathscr{B}}_{T_N}} \frac{1}{N} \sum_{i=1}^{N} \|w^i - \widehat{w}^i\|_2^2 \tag{M}$$

of a given model $\widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$. The problem decouples into N independent sub-problems

$$M(\mathscr{B}, \mathscr{D}) = \sum_{i=1}^{N} M(\mathscr{B}, w^i),$$

where

$$M(\mathscr{B}, w) = \min_{\widehat{w} \in \widehat{\mathscr{B}}_T} \frac{1}{N} \|w - \widehat{w}\|_2^2,$$

which are classical least-norm problems. Their solution is given by

$$M(\mathscr{B}, w) = \|\Pi_{\mathscr{B}_T}^{\perp} w\|_2^2,$$

where $\Pi_{\mathscr{B}_T}^{\perp}$ is the projector on the orthogonal complement $\mathscr{B}_T^{\perp}$ of the subspace $\mathscr{B}_T$.

The elimination of $\widehat{w}^1, \ldots, \widehat{w}^N$ from (ML) results in the equivalent problem

$$\text{minimize} \quad M(\widehat{\mathscr{B}}, \mathscr{D}) \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}, \tag{ML'}$$

where

$$M(\widehat{\mathscr{B}}, \mathscr{D}) = \frac{1}{N} \sum_{i=1}^{N} \|\Pi_{\mathscr{B}_{T_i}}^{\perp} w^i\|_2^2.$$

V. Consistency of the maximum likelihood estimator

Theorem 4 (Consistency)**.** *Assuming that the data $\mathscr{D}$ is generated in the errors-in-variables setup (EIV) and (A) holds true, the estimator defined by (ML') is strongly consistent,* i.e.,

$$\widehat{\mathscr{B}} \to \overline{\mathscr{B}} \text{ with probability 1 as } N \to \infty.$$

Proof. Under the assumptions of the errors-in-variables model, the expected value of $M(\widehat{\mathscr{B}}, w^i)$ is

$$\begin{aligned} \mathbf{E}\left(M(\widehat{\mathscr{B}}, w^i)\right) &= \mathbf{E}(\overline{w}^i + \widetilde{w}^i)^\top \Pi^\perp_{\mathscr{B}_T}(\overline{w}^i + \widetilde{w}^i) \\ &= (\overline{w}^i)^\top \Pi^\perp_{\mathscr{B}_T} \overline{w}^i + \mathbf{E}\left((\widetilde{w}^i)^\top \Pi^\perp_{\mathscr{B}_{T_i}} \widetilde{w}^i\right). \end{aligned} \tag{$*$}$$

Using the fact that $\widetilde{w}$ is white noise with covariance $\sigma^2 I$ and $\Pi^\perp_{\mathscr{B}_T}$ is a projector, we have for the second term

$$\mathbf{E}\left((\widetilde{w}^i)^\top \Pi^\perp_{\mathscr{B}_{T_i}} \widetilde{w}^i\right) = \sigma^2 \operatorname{trace}(\Pi^\perp_{\mathscr{B}_{T_i}}) = \sigma^2 \big(\mathtt{m}T_i + \ell(q - \mathtt{m})\big).$$

Therefore, the second term of $(*)$ does not depend on the model $\widehat{\mathscr{B}}$. The first term of $(*)$ is minimized for $\widehat{\mathscr{B}} = \overline{\mathscr{B}}$. Indeed, by definition of the projection matrix

$$(\overline{w}^i)^\top \Pi^\perp_{\overline{\mathscr{B}}_{T_i}} \overline{w}^i = 0.$$

We have shown that $\overline{\mathscr{B}}$ is a global minimizer of $\mathbf{E}M$. By the strong law of large numbers [18],

$$\lim_{N\to\infty} M(\widehat{\mathscr{B}}) \to \mathbf{E}M(\widehat{\mathscr{B}}) \text{ with probability 1.}$$

The limit is finite and nonzero by (A). Then, by [19], the minimizer of M, *i.e.*, the estimator $\widehat{\mathscr{B}}$ converges with probability 1 to the minimizer of $\mathbf{E}M$, which is $\overline{\mathscr{B}}$. □

The result of Theorem 4 is illustrated next in the case of an autonomous linear time-invariant system, parameterized by their poles, *i.e.*, 1) $\mathtt{m} = 0$ and 2) a specific representation of the system is chosen. See [20] for an application of this special case in operational modal analysis.

VI. Autonomous systems

A scalar autonomous linear time-invariant system with simple poles can be represented by the "sum-of-exponentials model" [21], [22]

$$\mathscr{B} = \left\{ y = \sum_{j=1}^{\mathtt{n}} c_j \exp_{z_j} \mid c \in \mathbb{C}^{\mathtt{n}} \right\}. \tag{SDE}$$

Here $\exp_z$ is the exponential function $\exp_z(t) := z^t$. The complex numbers $z_1, \ldots, z_{\mathtt{n}}$ are the *poles* of the system. In the representation (SDE), they are assumed to be distinct, *i.e.*, $z_i \neq z_j$, for all $i \neq j$.

A finite trajectory $y = \big(y(1), \ldots, y(T)\big)$ of a sum-of-exponentials model (SDE) can be expressed as

$$y = P_T(\theta)c,$$

where $P_T(\theta)$ is the (extended) Vandermonde matrix

$$P_T(\theta) := \begin{bmatrix} 1 & \cdots & 1 \\ z_1 & \cdots & z_{\mathrm{n}} \\ \vdots & & \vdots \\ z_1^{T-1} & \cdots & z_{\mathrm{n}}^{T-1} \end{bmatrix}$$

and $z_1, \ldots, z_{\mathrm{n}}$ are the roots of the polynomial

$$\theta(z) := \theta_1 + \theta_2 z + \cdots + \theta_{\mathrm{n}} z^{\mathrm{n}-1} + z^{\mathrm{n}}.$$

In the autonomous case, the errors-in-variables setup (EIV) coincides with the output error setup

$$\begin{aligned} y^i = \overline{y}^i + \widetilde{y}^i, \quad & \text{where } \overline{y}^i \in \overline{\mathscr{B}} \in \mathscr{L}_{0,\mathrm{n}} \\ & \widetilde{y}^i \text{ is zero mean white Gaussian} \\ & \text{process with covariance } s^2 I \\ & \text{and } \widetilde{y}^i \text{ is independent of } \widetilde{y}^j \\ & \text{for all } i \neq j. \end{aligned} \tag{OE}$$

The maximum-likelihood identification problem (ML) specialized for the sum-of-exponentials model becomes

$$\begin{aligned} &\text{minimize} \quad \frac{1}{N} \sum_{i=1}^{N} \|y^i - \widehat{y}^i\|_2^2 \quad \text{over } \theta,\ c^1, \ldots, c^N,\ \widehat{y}^1, \ldots, \widehat{y}^N \\ &\text{subject to} \quad \widehat{y}^i = P_{T_i}(\theta) c^i, \quad \text{for } i = 1, \ldots, N \end{aligned}$$

or

$$\text{minimize} \quad \frac{1}{N} \sum_{i=1}^{N} \|y^i - P_{T_i}(\theta) c^i\|_2^2 \quad \text{over } \theta,\ c^1, \ldots, c^N. \tag{ML$_{\text{aut}}$}$$

Applying the variable projections method to (ML$_{\text{aut}}$), leads to N decoupled problems

$$\text{minimize} \quad \frac{1}{N} \|y^i - P_{T_i}(\theta) c^i\|_2^2 \quad \text{over } c^i, \text{ for } i = 1, \ldots, N.$$

These are ordinary least squares problems with solutions

$$\widehat{y}^i = \underbrace{P_{T_i}(\theta)\big(P_{T_i}^\top(\theta) P_{T_i}(\theta)\big)^{-1} P_{T_i}}_{\Pi_{T_i}(\theta)} y^i,$$

where $\Pi_{T_i}(\theta)$ is an idempotent matrix ($\Pi_{T_i}^2(\theta) = \Pi_{T_i}(\theta)$) Therefore, the cost function of the sum-of-exponentials model is

$$M(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y^i)^\top \big(I - \Pi_{T_i}(\theta)\big) y^i.$$

Proposition 5 (Consistency in the autonomous case with (SDE) representation)**.** *Assuming that the data $\mathscr{D}$ is generated in the output error setup (OE), the estimator defined by (ML_{aut}) is strongly consistent,* i.e.,

$$\widehat{\theta} \to \overline{\theta} \text{ with probability 1 as } N \to \infty.$$

Proposition 5 is a corollary of Theorem 4, however, since the aim of this section is to illustrate how the general result specializes in the autonomous case and by choosing a particular model representation, an independent proof of Proposition 5 is given in the Appendix.

VII. Numerical example

As mentioned in the introduction, there is readily available software [6], [11] for identification of linear time-invariant systems from multiple experiments. In this section, we use the software of [6], [11] to illustrate the consistency and asymptotic efficiency properties of the maximum likelihood estimator (ML) on a simulation example.

A. Consistency

The data generating system is an autonomous continuous-time linear time-invariant system of order $\mathtt{n} = 6$ with with resonance angular frequencies

$$\bar{\omega}_1 = 2\pi 80 \,\text{rad/s}, \;\; \bar{\omega}_2 = 2\pi 130 \,\text{rad/s}, \;\; \bar{\omega}_3 = 2\pi 200 \,\text{rad/s}$$

and poles' damping ratios

$$\zeta_1 = 0.15, \quad \zeta_2 = 0.1, \quad \zeta_3 = 0.2.$$

The system is sampled with a period $t_s = 10^{-3}$. A specific response $\bar{y}$ is shown in Figure 1.

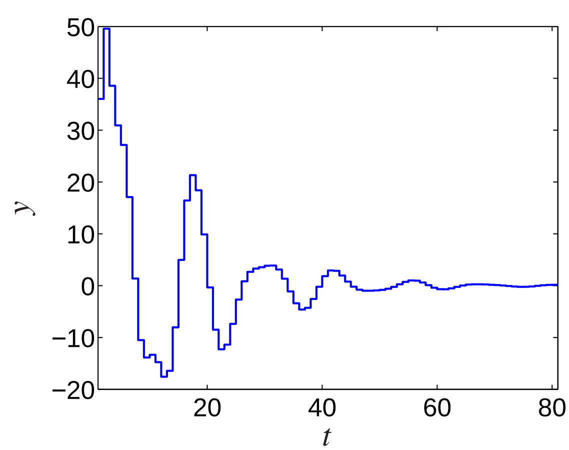

Fig. 1. Specific trajectory of $\overline{\mathscr{B}}$.

The identification data $\mathscr{D}$ is generated via the output error model (OE) with random true trajectories $\overline{y}^1,\ldots,\overline{y}^N$. The lengths $T_1,\ldots,T_N$ of the responses are determined, so that the trajectories have sufficient decay.

The number of experiments N varies from 1 to 50 and the signal-to-noise ratio is 100. For each value of N, the identification experiment is repeated $K = 200$ times and the average parameter error

$$e = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\|\overline{\theta}-\widehat{\theta}^k\|_2^2}.$$

is computed. The results obtained (see Figure 2) show the convergence of the average parameter estimation error. The convergence rate is close to $1/\sqrt{N}$.

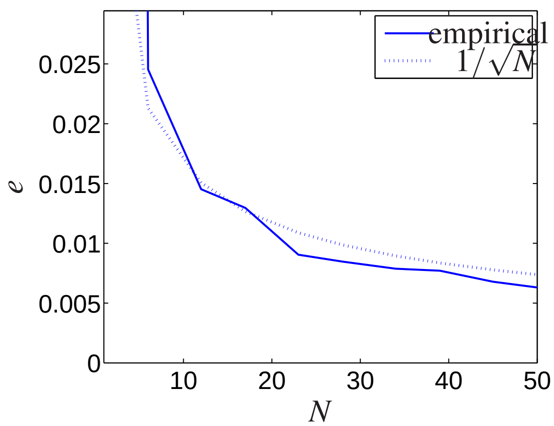

Fig. 2. Error convergence.

B. *Efficiency*

In this section, we compare empirically the maximum-likelihood estimate covariance matrix $V_{\widehat{\theta}} = \operatorname{cov}(\widehat{\theta})$ and the Cramér-Rao lower bound C, see [2, Chapter 19]. Both are evaluated for an experiment with N data sets, where N is increased from 1 to 50. In the simulation setup of the example in Section VII-A the criterion

$$e' = \frac{\|C - V_{\widehat{\theta}}\|_2}{\|C\|_2},$$

where $\|\cdot\|_2$ is the spectral norm, is shown in Figure 3 shows as a function of N.

The result suggests that the maximum-likelihood estimate covariance matrix does not converge to the Cramér-Rao lower bound, however, for high signal-to-noise ratio the discrepancy is small. This empirical observation can be shown analytically. The question of whether the maximum likelihood estimator is asymptotically inefficiency or the Cramér-Rao lower bound is conservative remains open.

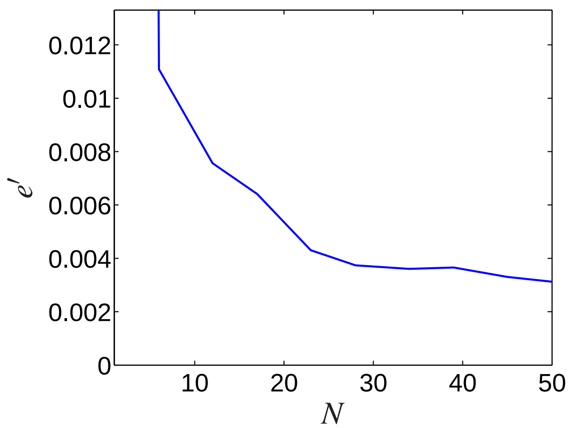

Fig. 3. Relative error between the maximum-likelihood estimate covariance matrix and the Cramér-Rao lower bound.

APPENDIX

Under the data generation assumptions (OE), the expected value of the cost function is

$$\begin{aligned}\mathbf{E}M(\theta) &= \frac{1}{N}\mathbf{E}\sum_{i=1}^{N}\|\big(I-\Pi_{T_i}(\theta)\big)(\overline{y}^i+\widetilde{y}^i)\|_2^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\|\big(I-\Pi_{T_i}(\theta)\big)\overline{y}^i\|_2^2 \\ &\qquad + \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\|\big(I-\Pi_{T_i}(\theta)\big)\widetilde{y}^i)\|_2^2.\end{aligned}$$

For the second term we have

$$\begin{aligned}\mathbf{E}\|\big(I-\Pi_{T_i}(\theta)\big)\widetilde{y}^i\|_2^2 &= \mathbf{E}\Big((\widetilde{y}^i)^\top\big(I-\Pi_{T_i}(\theta)\big)\widetilde{y}^i\Big) \\ &= \sum_{j=1}^{T_i}\sum_{k=1}^{T_i}\big(\delta_{jk}-\Pi_{T_i,jk}(\theta)\big)\,\mathbf{E}(\widetilde{y}^i_j\widetilde{y}^i_k),\end{aligned}$$

where $\delta_{jk}=1$ if $j=k$ and 0 otherwise. By (OE), $\mathbf{E}(\widetilde{y}^i_j\widetilde{y}^i_k)=s^2 I_{\mathtt{p}}\delta_{jk}$, so that

$$\mathbf{E}\|I=\Pi_{T_i}(\theta)\|_2^2 = \operatorname{trace}\big(I-\Pi_{T_i}(\theta)\big)s^2.$$

Since $I-\Pi_{T_i}(\theta)$ is a projector matrix of rank $T_i\mathtt{p}-\mathtt{n}$, its eigenvalues are $T_i\mathtt{p}-\mathtt{n}$ ones and $\mathtt{n}$ zeros. Therefore,

$$\mathbf{E}\|\big(I-\Pi_{T_i}(\theta)\big)\widetilde{y}^i\|_2^2 = (T_i\mathtt{p}-\mathtt{n})s^2$$

and is independent of θ.

For the true parameter vector $\overline{\theta}$ we have that $M(\overline{\theta})=0$, so that $\overline{\theta}$ is a global minimizer of $\mathbf{E}M$. By the strong law of large numbers,

$$\lim_{N\to\infty} M(\theta) \to \mathbf{E}M(\theta) \text{ with probability } 1.$$

The limit is finite and nonzero by (A). Then, by [19], the minimizer of M, *i.e.*, the estimator $\widehat{\theta}$ converges with probability 1 to the minimizer of $\mathbf{E}M$, which is $\overline{\theta}$.

Acknowledgements

The motivation to study the consistency properties of the maximum likelihood estimator in the autonomous case comes from joint work with Jan Goos and Konstantin Usevich [23].

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement number 258581 "Structured low-rank approximation: Theory, algorithms, and applications", the Research Foundation Flanders (FWO-Vlaanderen), the Flemish Government (Methusalem Fund, METH1), and the Belgian Federal Government (Interuniversity Attraction Poles programme VII, Dynamical Systems, Control, and Optimization).

References

[1] J. C. Willems, P. Rapisarda, I. Markovsky, and B. D. Moor, "A note on persistency of excitation," *Control Lett.*, vol. 54, no. 4, pp. 325–329, 2005.

[2] J. Schoukens, G. Vandersteen, Y. Rolain, and R. Pintelon, "Frequency response function measurements using concatenated subrecords with arbitrary length," *IEEE Trans. on Instr. and Measurement*, vol. 61, no. 10, pp. 2682–2688, 2012.

[3] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[4] T. Söderström and P. Stoica, *System Identification*. Prentice Hall, 1989.

[5] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*, 2nd ed. Piscataway, NJ: IEEE Press, 2012.

[6] I. Markovsky, "A software package for system identification in the behavioral setting," *Control Engineering Practice*, vol. 21, pp. 1422–1436, 2013.

[7] D. Tufts and A. Shah, "Estimation of a signal waveform from noisy data using low-rank approximation to a data matrix," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1716–1721, 1993.

[8] B. De Moor, "Total least squares for affinely structured matrices and the noisy realization problem," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3104–3113, 1994.

[9] I. Markovsky, "Structured low-rank approximation and its applications," *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

[10] K. Usevich and I. Markovsky, "Variable projection for affinely structured low-rank approximation in weighted 2-norms," *J. Comput. Appl. Math.*, vol. 272, pp. 430–448, 2014.

[11] I. Markovsky and K. Usevich, "Software for weighted structured low-rank approximation," *J. Comput. Appl. Math.*, vol. 256, pp. 278–292, 2014.

[12] G. Golub and V. Pereyra, "Separable nonlinear least squares: the variable projection method and its applications," *Institute of Physics, Inverse Problems*, vol. 19, pp. 1–26, 2003.

[13] P. Guillaume and R. Pintelon, "A Gauss–Newton-like optimization algorithm for "weighted" nonlinear least-squares problems," *IEEE Trans. Signal Process.*, vol. 44, no. 9, pp. 2222–2228, 1996.

[14] J. C. Willems, "From time series to linear system—Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling," *Automatica*, vol. 22, 23, pp. 561–580, 675–694, 87–115, 1986, 1987.

[15] J. Polderman and J. C. Willems, *Introduction to mathematical systems theory*. Springer-Verlag, 1998.

[16] C. Heij and W. Scherrer, "Consistency of system identification by global total least squares," *Automatica*, vol. 35, pp. 993–1008, 1999.

[17] T. Georgiou and M. Smith, "Optimal robustness in the gap metric," *IEEE Trans. Automat. Contr.*, vol. 35, no. 6, pp. 673–686, 1990.

[18] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*. Springer, 1988.

[19] T. Söderström, "Convergence properties of the generalised least squares identitication method," *Automatica*, vol. 10, no. 6, pp. 617–626, 1974.

[20] I. Markovsky and R. Pintelon, "Consistent estimation of autonomous linear time-invariant systems from multiple experiments," in *Proc. of the Conference on Noise and Vibration Engineering (ISMA)*, Leuven, Belgium, September 2014.

[21] R. Kumaresan and D. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Trans. Acust., Speech, Signal Process.*, vol. 30, no. 6, pp. 833–840, 1982.

[22] T. Abatzoglou, J. Mendel, and G. Harada, "The constrained total least squares technique and its application to harmonic superresolution," *IEEE Trans. Signal Process.*, vol. 39, pp. 1070–1087, 1991.

[23] I. Markovsky, J. Goos, K. Usevich, and R. Pintelon, "Realization and identification of autonomous linear periodically time-varying systems," *Automatica*, vol. 50, pp. 1632–1640, 2014.