System identification in the behavioral setting
A structured low-rank approximation approach

Ivan Markovsky

Department ELEC
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, Building K, B-1050
Brussels, Belgium

imarkovs@vub.ac.be

Abstract. System identification is a fast growing research area that encompasses a broad range of problems and solution methods. It is desirable to have a unifying setting and a few common principles that are sufficient to understand the currently existing identification methods. The behavioral approach to system and control, put forward in the mid 80's, is such a unifying setting. Till recently, however, the behavioral approach lacked supporting numerical solution methods. In the last 10 yeas, the structured low-rank approximation setting was used to fulfill this gap. In this paper, we summarize recent progress on methods for system identification in the behavioral setting and pose some open problems. First, we show that errors-in-variables and output error system identification problems are equivalent to Hankel structured low-rank approximation. Then, we outline three generic solution approaches: 1) methods based on local optimization, 2) methods based on convex relaxations, and 3) subspace methods. A specific example of a subspace identification method—data-driven impulse response computation—is presented in full details. In order to achieve the desired unification, the classical ARMAX identification problem should also be formulated as a structured low-rank approximation problem. This is an outstanding open problem.

Keywords: system identification; errors-in-variables modeling, behavioral approach; Hankel matrix, low-rank approximation, impulse response estimation, ARMAX identification.

1 Introduction

System identification aims at deriving a dynamical model $\widehat{\mathscr{B}}$ (*i.e.*, a mathematical description) from observed data $\mathscr{D}$ of a to-be-modeled physical plant. The data is typically obtained by sampling and quantization in time-domain from one or more independent measurement experiments. Each measurement point is a real-valued vector of the observed variables from the system and the model postulates a relation among the variables.

Prior knowledge and/or assumptions about the plant are incorporated in the identification problem by restricting the model to belong to a set of models $\mathscr{M}$, called the

model class. Therefore, a system identification problem is a mapping:

$$\begin{array}{c}\text{data}\\ \mathscr{D}\end{array} \quad \xrightarrow{\begin{array}{c}\text{system}\\ \text{identification}\end{array}} \quad \begin{array}{c}\text{model}\\ \widehat{\mathscr{B}} \in \mathscr{M}\end{array} \tag{ID}$$

The mapping (ID) is defined implicitly as a solution to an optimization problem, *i.e.*, the model $\widehat{\mathscr{B}}$ minimizes (among all feasible models) a specified cost function. Different identification problems correspond to different choices of a model class and the cost function.

Two contradictory objectives in system identification are:

1. "simple" model,
2. "good" fit of the data by the model.

Typically the model class is used to impose a hard bound on the model complexity and the cost function is used to measure the model-data misfit (lack of fit). It is possible, however, to minimize the model complexity subject to a hard bound on the misfit or, more generally, consider the bi-objective minimization of the complexity and the misfit.

In exact identification [10, Ch. 7], the model complexity is minimized subject to the constraint that the model fits the data exactly (zero misfit). If such a model exists in the model class, it is called the *most powerful unfalsified model* (of $\mathscr{D}$ in $\mathscr{M}$) [13]. Exact identification is a theoretical tool which is a generalization of the realization problem in system theory and appears in approximate and stochastic identification problems [7].

The data collected in a real-life experiment is "inexact" due to disturbances (unobserved variables), measurement noises, discretization, and quantization errors. Methods for computing the most powerful unfalsified model, however, lead through simple modifications to a class of practical identification methods known as subspace methods.

In this paper, we consider the model class of linear time-invariant systems of bounded complexity $\mathscr{L}_{\mathtt{m},\ell}$ — number of input variables at most $\mathtt{m}$ and lag at most ℓ. In the behavioral setting [14], no a priori separation of the variables into inputs and outputs is made, however any model allows a nonunique input/output partition. Although the choice of the input variables is in general not unique, the number of inputs is a model invariant, *i.e.*, it does not depend on the partitioning.

In Section 2, we define the model class $\mathscr{L}_{\mathtt{m},\ell}$ and the approximation criterion, which specify the identification problem (ID). The misfit has the geometric interpretation of the Euclidean distance between the data and the model. In the stochastic setting, misfit minimization corresponds to errors-invariables system identification [12], *i.e.*, (ID) is a maximum-likelihood estimator in the errors-invariables setting. Section 3 related the identification problem (ID) to the weighted Hankel structured low-rank approximation. Three generic classes of solution methods are outlined: local optimization based methods, convex relaxation based methods, and subspace methods. As a specific example of a subspace method, in Appendix A, we present a data-driven algorithm for impulse response estimation. Section 4 draws conclusions and states some open problems. One of them is integration of the classical ARMAX setting in the behavioral setting.

2 Problem formulation

A dynamical system $\mathscr{B}$ is a set of trajectories. In discrete-time, a trajectory is a q-variable time-series $w : \mathbb{Z} \to \mathbb{R}^q$. The class of finite dimensional linear time-invariant systems with at most $\mathtt{m}$ inputs is denoted by $\mathscr{L}_{\mathtt{m}}$. This class admits a representation

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

where σ is the shift operator $(\sigma w)(t) = w(t+1)$. The smallest number ℓ, for which there is ℓth order representation $\mathscr{B} = \mathscr{B}(R)$ is called the *lag* of the system. The pair $(\mathtt{m}, \ell)$ specify the model complexity. The model class of bounded complexity is denoted by $\mathscr{L}_{\mathtt{m},\ell}$.

The model variables w can be partitioned into inputs u and outputs y, *i.e.*, there is a permutation matrix Π, such that $w = \Pi \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$. The system can then be represented in the classical form

$$\mathscr{B} = \mathscr{B}(A,B,C,D,\Pi) := \{\, w = \Pi \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right] \mid \text{there is } x, \text{ such that } \sigma x = Ax + Bu,\ y = Cx + Du \,\}. \tag{I/S/O}$$

We will assume that Π can be chosen equal to I and the block $P_\ell \in \mathbb{R}^{\mathtt{p}\times\mathtt{p}}$ of $R_\ell = \begin{bmatrix} Q_\ell & -P_\ell \end{bmatrix}$ in a difference equation representation is nonsingular.

Let the identification data $\mathscr{D}$ be an observed trajectory

$$w_{\mathrm{d}} = \big(w_{\mathrm{d}}(1), \ldots, w_{\mathrm{d}}(T)\big), \quad w_{\mathrm{d}}(t) \in \mathbb{R}^q$$

of the to-be-identified system. The approximation criterion, called data-model misfit, is defined as follows:

$$M(\mathscr{D}, \mathscr{B}) := \min_{\widehat{w} \in \mathscr{B}} \|w_{\mathrm{d}} - \widehat{w}\|_2, \tag{M}$$

where $\widehat{w}$ is the optimal approximation of w_{d} by $\mathscr{B}$. Note that $\widehat{w}$ is the projection of w_{d} on $\mathscr{B}$.

The identification problem considered is misfit minimization over all system $\widehat{\mathscr{B}}$ in the model class $\mathscr{L}_{\mathtt{m},\ell}$:

$$\text{minimize} \quad \text{over } \mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell} \quad M(w_{\mathrm{d}}, \mathscr{B}). \tag{SYSID}$$

Generalizations of problem (SYSID) (see [8]) are weighted 2-norm approximation criteria, specification of exact and missing variables, and data consisting on multiple trajectories.

3 Hankel low-rank approximation

In what follows, we will use the block-Hankel matrix

$$\mathscr{H}_{\ell+1}(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(T-\ell) \\ w(2) & w(3) & \cdots & w(T-\ell+1) \\ \vdots & \vdots & & \vdots \\ w(\ell+1) & w(\ell+2) & \cdots & w(T) \end{bmatrix}.$$

The fundamental link between the system identification problem (SYSID) and structured low-rank approximation is the following equivalence

$$w \in \mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell} \quad \Longleftrightarrow \quad \operatorname{rank}\big(\mathscr{H}_{\ell+1}(w)\big) \leq (\ell+1)\mathtt{m} + \mathtt{p}\ell. \tag{$*$}$$

In words, w is an (exact) trajectory of the linear time-invariant system $\mathscr{B}$ if and only if the Hankel structured matrix $\mathscr{H}_{\ell+1}(w_{\mathrm{d}})$ is rank deficient. Note that the complexity of the model $\mathscr{B}$ (number of inputs $\mathtt{m}$ and lag ℓ) is directly related to the rank constraint of the Hankel matrix.

Using $(*)$, we can rewrite the identification problem (SYSID) as an equivalent Hankel low-rank approximation problem

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{w} \quad \|w - \widehat{w}\|^2 \\ \text{subject to} & \operatorname{rank}\big(\mathscr{H}_{\ell+1}(\widehat{w})\big) \leq (\ell+1)\mathtt{m} + \mathtt{p}\ell. \end{array} \tag{SLRA}$$

The main issue in system identification is that Problem (SLRA) is nonconvex. Therefore, various heuristics, reviewed later, are used for its solution.

3.1 Alternating projections approach

One approach for dealing with the rank constraint in (SLRA) is to use the kernel representation

$$\begin{aligned} \operatorname{rank}\big(\mathscr{H}_{\ell+1}(\widehat{w})\big) \leq r \quad \Longleftrightarrow \quad & R\mathscr{H}_{\ell+1}(\widehat{w}) = 0 \\ & \text{and} \quad R \in \mathbb{R}^{\mathtt{p}\times(\ell+1)q} \text{ is full row rank.} \end{aligned} \tag{KER}$$

Using (KER), (SLRA) becomes a classical parameter optimization problem,

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{w} \text{ and } R \quad \|w_{\mathrm{d}} - \widehat{w}\|^2 \\ \text{subject to} & R\mathscr{H}_{\ell+1}(\widehat{w}) = 0 \quad \text{and} \quad R \text{ is f.r.r.} \end{array} \tag{SLRA$_R$}$$

(SLRA$_R$) is furthermore equivalent to

$$\text{minimize} \quad \text{over f.r.r. } R \in \mathbb{R}^{(m-r)\times m} \quad M(R), \tag{OUTER}$$

where

$$\begin{array}{ll} M(R) := \min\limits_{\widehat{w}} & \|w_{\mathrm{d}} - \widehat{w}\|^2 \\ \text{subject to} & R\mathscr{H}_{\ell+1}(\widehat{w}) = 0. \end{array} \tag{INNER}$$

Note that (INNER) is a classical linear least squares problem.

The approach for solving (SLRA$_R$) by minimizing (OUTER) is closely related to the variable projection method in numerical linear algebra [4]. In [4], however, an explicit function $\widehat{b} = A(\theta)x$, where x is unconstrained, is considered, while in the context of the structured low-rank approximation problem, an implicit relation $R\mathscr{H}_{\ell+1}(\widehat{w}) = 0$ is considered, where the variable R is constrained to have full row rank. This fact requires

new type of algorithms where the nonlinear least squares problem is an optimization problem on a Grassmann manifold, see [1, 2].

In (OUTER), the cost function M is minimized over the set of full row rank matrices R. Indeed, M depends only on the space spanned by the rows of R. In order to find a minimum of M, the search space in (OUTER) can be replaced by the matrices satisfying the constraint

$$RR^\top = I_{\mathrm{p}}.$$

A software package for Hankel structured low-rank approximation is presented in [9]. The Levenberg-Marquardt algorithm [11] implemented the GNU Scientific Library [3], are used for the solution of the nonlinear least squares problem. This package is used in [8] for system identification.

3.2 Alternating projections approach

The second approach is based on the image representation of the rank constraint

$$\operatorname{rank}\left(\mathscr{H}_{\ell+1}(\widehat{w})\right) \leq r \quad \Longleftrightarrow \quad \mathscr{H}_{\ell+1}(\widehat{w}) = PL \quad \text{where} \quad P \in \mathbb{R}^{\bullet\times r} \text{ and } L \in \mathbb{R}^{r\times\bullet} \qquad \text{(KER)}$$

and a representation of a structured matrix by an linear equality constraint

$$\Pi(PL) - PL = 0,$$

where Π is a projection of a matrix to the nearest one with Hankel structure.

3.3 Nuclear norm heuristic

The nuclear norm heuristic replaces the rank constraint in (SLRA) with a constraint $\|\mathscr{H}_{\ell+1}(\widehat{w})\|_* \leq \gamma$ on the nuclear norm $\|\cdot\|_*$ of $\mathscr{H}_{\ell+1}(\widehat{w})$. The nuclear norm is a convex function of $\widehat{w}$, so that the relaxed problem is a convex optimization problem. The parameter γ is selected in [6] by bisection aiming at achievement of the desired rank of the approximation $\mathscr{H}_{\ell+1}(\widehat{w})$. Other authors [5], however, select a value of γ that does not necessarily lead to rank deficient structured matrix. In this case, the nuclear norm minimization is used as a data preprocessing step. The obtained $\widehat{w}$ from the nuclear norm minimization is then fed as an input to a subspace method, which does the actual rank reduction.

3.4 Subspace methods

The subspace methods for approximate system identification originate from corresponding methods for exact system identification, by replacing exact operations such as rank revealing factorization and solution of a system of linear equations by approximate methods — unstructured low-rank approximation (achieved via the singular value decomposition) and approximation solution of a system of linear equations in the least squares sense. Since two or more steps of the algorithm are adapted in this way, the result heuristic methods can be called multi-stage methods. They are suboptimal, however, they are fast and effective methods for approximate system identification. A detailed specific example of a subspace method is shown in Appendix A.

4 Conclusions and future perspectives

In this paper, we described a unifying setting for system identification as a biobjective optimization problem. The identified model is defined in the behavioral sense as a set of trajectories. The two objectives are 1) minimization of the fitting error and 2) minimization of the model complexity. As a specific example of a fitting error, we gave the misfit, *i.e.*, the projection of the data on the model. This error criterion corresponds to the class of the errors-in-variables problems in the system identification literature. Another error criterion is the latency, which corresponds to the class of the ARMAX identification problems.

The main computation tool in the behavioral setting is Hankel structured low-rank approximation. The link to low-rank approximation follows from the fact that a time series is a trajectory of a linear time invariant system if and only if a Hankel structured matrix composed of the data is rank deficient. Once the identification problem is re-formulated as a structured low-rank approximation problem, it can be solved by various methods. The methods however are classified into three groups: local optimization based methods, convex relaxations, and subspace methods. In general, the subspace methods are faster but less efficient than the optimization based methods.

The class of methods based on convex relaxations are currently actively developed. The main challenges in this area of research are finding theoretical bounds for the distance to global optimality and development of efficient computational methods.

Another research challenge is formulation of the latency minimization (ARMAX system identification) as a structured low-rank approximation and solution of the resulting problem by existing methods for low-rank approximation.

Acknowledgements

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".

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008)
2. Absil, P.A., Mahony, R., Sepulchre, R., Dooren, P.V.: A Grassmann–Rayleigh quotient iteration for computing invariant subspaces. SIAM Review 44(1), 57–73 (2002)
3. Galassi, M., et al.: GNU scientific library reference manual. `http://www.gnu.org/software/gsl/`
4. Golub, G., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. Institute of Physics, Inverse Problems 19, 1–26 (2003)
5. Liu, Z., Hansson, A., Vandenberghe, L.: Nuclear norm system identification with missing inputs and outputs. Control Lett. 62, 605–612 (2013)

6. Markovsky, I.: How effective is the nuclear norm heuristic in solving data approximation problems? In: Proc. of the 16th IFAC Symposium on System Identification. pp. 316–321. Brussels (2012)
7. Markovsky, I.: Low Rank Approximation: Algorithms, Implementation, Applications. Springer (2012)
8. Markovsky, I.: A software package for system identification in the behavioral setting. Control Engineering Practice 21, 1422–1436 (2013),
9. Markovsky, I., Usevich, K.: Software for weighted structured low-rank approximation. J. Comput. Appl. Math. 256, 278–292 (2014),
10. Markovsky, I., Willems, J.C., Van Huffel, S., De Moor, B.: Exact and Approximate Modeling of Linear Systems: A Behavioral Approach. No. 11 in Monographs on Mathematical Modeling and Computation, SIAM (March 2006),
11. Marquardt, D.: An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. 11, 431–441 (1963)
12. Söderström, T.: Errors-in-variables methods in system identification. Automatica 43, 939–958 (2007)
13. Willems, J.C.: From time series to linear system—Part II. Exact modelling. Automatica 22(6), 675–694 (1986)
14. Willems, J.C.: From time series to linear system—Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling. Automatica 22, 23, 561–580, 675–694, 87–115 (1986, 1987)
15. Willems, J.C., Rapisarda, P., Markovsky, I., Moor, B.D.: A note on persistency of excitation. Control Lett. 54(4), 325–329 (2005)

A Subspace method for impulse response estimation

Let $\mathscr{B}$ be a linear time-invariant system of order $\mathtt{n}$ with lag ℓ and let $w = (u,y)$ be an input-output partitioning of the variables. In [15], it is shown that, under the following conditions,

- the data w_d is exact, *i.e.*, $w_\mathrm{d} \in \mathscr{B}$,
- $\mathscr{B}$ is controllable,
- u_d is persistently exciting, *i.e.*, $\mathscr{H}_{\mathtt{n}+\ell+1}(u_\mathrm{d})$ is full rank,

the Hankel matrix $\mathscr{H}_t(w_\mathrm{d})$ with t block-rows, composed from w_d, spans the space $\mathscr{B}|_{[1,t]}$ of all t-samples long trajectories of the system $\mathscr{B}$, *i.e.*,

$$\operatorname{image}\big(\mathscr{H}_t(w_\mathrm{d})\big) = \mathscr{B}|_{[1,t]}.$$

This implies that there exists a matrix G, such that

$$\mathscr{H}_t(y_\mathrm{d})G = H,$$

where H is the vector of the first t samples of the impulse response of $\mathscr{B}$. The problem of computing the impulse response H from the data w_d reduces to the one of finding a particular G.

Define U_p, U_f, Y_p, Y_f as follows

$$\mathscr{H}_{\ell+t}(u_\mathrm{d}) =: \begin{bmatrix} U_\mathrm{p} \\ U_\mathrm{f} \end{bmatrix}, \qquad \mathscr{H}_{\ell+t}(y_\mathrm{d}) =: \begin{bmatrix} Y_\mathrm{p} \\ Y_\mathrm{f} \end{bmatrix},$$

where

$$\operatorname{row}\dim(U_{\mathrm{p}}) = \operatorname{row}\dim(Y_{\mathrm{p}}) = \ell$$

and

$$\operatorname{row}\dim(U_{\mathrm{f}}) = \operatorname{row}\dim(Y_{\mathrm{f}}) = t.$$

Then if $w_{\mathrm{d}} = (u_{\mathrm{d}}, y_{\mathrm{d}})$ is a trajectory of a controllable linear time-invariant system $\mathscr{B}$ of order $\mathtt{n}$ and lag ℓ and if u_{d} is persistently exciting of order $t + \ell + \mathtt{n}$, the system of equations

$$\begin{bmatrix} U_{\mathrm{p}} \\ U_{\mathrm{f}} \\ Y_{\mathrm{p}} \end{bmatrix} G = \begin{bmatrix} 0_{\mathtt{m}\ell \times \mathtt{m}} \\ \begin{bmatrix} I_{\mathtt{m}} \\ 0_{\mathtt{m}(t-1)\times \mathtt{m}} \end{bmatrix} \\ 0_{\mathtt{p}\ell \times \mathtt{m}} \end{bmatrix}, \tag{$*$}$$

is solvable for $G \in \mathbb{R}^{\bullet \times \mathtt{m}}$, and for any particular solution G, the matrix $Y_{\mathrm{f}}G$ contains the first t samples of the impulse response of $\mathscr{B}$, *i.e.*,

$$Y_{\mathrm{f}}G = H.$$

This gives Algorithm 1 for the computation of H.

Algorithm 1 Block computation of the impulse response from data.

Input: u_{d}, y_{d}, ℓ, and t.

1: Solve the system of equations $(*)$ and let G be the computed solution.
2: Compute $H = Y_{\mathrm{f}}G$.

Output: H.

Algorithm 1 computes the first t samples of the impulse response; however, the persistency of excitation condition imposes a limitation on how big t can be. This limitation can be avoided by a modification of the algorithm. L consecutive samples, where L is a user specified parameter that is small enough to allow the application of Algorithm 1, are computed iteratively. Then, provided the system is stable, by monitoring the decay of H in the course of the computations, gives a way to determine how many samples are needed to capture the transient behavior of the system.