Available online at www.sciencedirect.com

ScienceDirect

automatica

Automatica 44 (2008) 891–909

www.elsevier.com/locate/automatica

Survey paper

Structured low-rank approximation and its applications[☆]

Ivan Markovsky

School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK

Received 7 November 2006; received in revised form 2 August 2007; accepted 5 September 2007
Available online 26 December 2007

Abstract

Fitting data by a bounded complexity linear model is equivalent to low-rank approximation of a matrix constructed from the data. The data matrix being Hankel structured is equivalent to the existence of a linear time-invariant system that fits the data and the rank constraint is related to a bound on the model complexity. In the special case of fitting by a static model, the data matrix and its low-rank approximation are unstructured.

We outline applications in system theory (approximate realization, model reduction, output error, and errors-in-variables identification), signal processing (harmonic retrieval, sum-of-damped exponentials, and finite impulse response modeling), and computer algebra (approximate common divisor). Algorithms based on heuristics and local optimization methods are presented. Generalizations of the low-rank approximation problem result from different approximation criteria (e.g., weighted norm) and constraints on the data matrix (e.g., nonnegativity). Related problems are rank minimization and structured pseudospectra.

Keywords: Low-rank approximation; Total least squares; System identification; Errors-in-variables modeling; Behaviors

1. Introduction

Fitting linear models to data can be achieved, both conceptually and algorithmically, by solving a system of equations $AX = B$, where the matrices A and B are constructed from the given data and the matrix X parameterizes the model to be found. In this classical approach, the main tools are the least squares method and its variations—data least squares (Degroat & Dowling, 1991), total least squares (TLS) (Golub & Van Loan, 1980), structured TLS (De Moor, 1993), robust least squares (Chandrasekaran, Gu, & Sayed, 1998), etc. The least squares method and its variations are mainly motivated by their applications for data fitting, but they invariably consider solving approximately an overdetermined system of equations.

In this paper we show that a number of linear data fitting problems are equivalent to the abstract problem of approximating a matrix D constructed from the data by a low-rank matrix. Partitioning the data matrix into matrices $A \in \mathbb{R}^{N\times \mathtt{m}}$ and $B \in \mathbb{R}^{N\times \mathtt{p}}$ and solving approximately the system $AX = B$ is a way to achieve rank-$\mathtt{m}$ or less approximation. The converse implication, however, is not true, because $[A \;\; B]$ having rank-$\mathtt{m}$ or less does not imply the existence of X, such that $AX = B$. This lack of equivalence between the original low-rank approximation problem and the $AX = B$ problem motivates what is called nongeneric TLS problem (Van Huffel & Vandewalle, 1991), whose theory is more complicated than the one of the generic problem and is difficult to solve numerically.

Alternative approaches for achieving a low-rank approximation are to impose that the data matrix has

1. at least $\mathtt{p} := \operatorname{coldim}(B)$ dimensional nullspace, or
2. at most $\mathtt{m} := \operatorname{coldim}(A)$ dimensional column space.

Parameterizing the nullspace and the column space by sets of basis vectors, the alternative approaches are:

1. *kernel representation*: there is a full rank matrix $R \in \mathbb{R}^{\mathtt{p}\times(\mathtt{m}+\mathtt{p})}$, such that $[A \;\; B]R^{\top} = 0$, and
2. *image representation*: there are matrices $P \in \mathbb{R}^{(\mathtt{m}+\mathtt{p})\times \mathtt{m}}$ and $L \in \mathbb{R}^{\mathtt{m}\times N}$, such that $[A \;\; B]^{\top} = PL$.

[☆] This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Wolfgang Scherrer under the direction of Editor Manfred Morari.

E-mail address: im@ecs.soton.ac.uk.

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2007.09.011

The approaches using kernel and image representations are equivalent to the original low-rank approximation problem. Next we illustrate the use of $AX = B$, kernel, and image representations on the most simple data fitting problem—line fitting.

1.1. Line fitting example

Given a set of points $\{d_1, \ldots, d_N\} \subset \mathbb{R}^2$ in the plane, the aim of the line fitting problem is to find a line passing through the origin that "best" matches the given points. The classical approach for line fitting is to define $\operatorname{col}(a_i, b_i) := d_i$ and solve approximately the overdetermined system

$$\operatorname{col}(a_1, \ldots, a_N)x = \operatorname{col}(b_1, \ldots, b_N) \tag{1}$$

by the least squares method. Let x_{ls} be the least squares approximate solution to (1). Then the least squares fitting line is

$$\mathscr{B}_{\text{ls}} := \{d = \operatorname{col}(a, b) \in \mathbb{R}^2 \mid ax_{\text{ls}} = b\}. \tag{2}$$

Geometrically, $\mathscr{B}_{\text{ls}}$ minimizes the sum of the squared vertical distances from the data points to the fitting line.

The left plot in Fig. 1 shows a particular example with $N = 10$ data points. The data points $d_1, \ldots, d_{10}$ are the circles in the figure, the fit $\mathscr{B}_{\text{ls}}$ is the solid line, and the fitting errors $e := ax_{\text{ls}} - b$ are the dashed lines. Visually we expect the best fit to be the vertical axis, so minimizing vertical distances is not appropriate in this example.

Note that by solving (1), we treat the a_i (the first components of the d_i) differently from the b_i (the second components): b_i is assumed to be a *function* of a_i. This is an arbitrary choice; we can as well fit the data by solving approximately the system

$$\operatorname{col}(a_1, \ldots, a_N) = \operatorname{col}(b_1, \ldots, b_N)x, \tag{3}$$

in which case a_i is assumed to be a function of b_i. Let x'_{ls} be the least squares approximate solution to (3). It gives the fitting line

$$\mathscr{B}'_{\text{ls}} := \{d = \operatorname{col}(a, b) \in \mathbb{R}^2 \mid a = bx'_{\text{ls}}\}, \tag{4}$$

which minimizes the sum of the squared horizontal distances (see the right plot in Fig. 1). The line $\mathscr{B}'_{\text{ls}}$ happens to achieve the desired fit in the example. This shows that

> in the classical method for data fitting, i.e., solving approximately a linear system of equations in the least squares sense, the choice of the model representation determines the fitting criterion.

This feature of the classical method is undesirable: it is more natural for a user of a data fitting method to specify a desired fitting criterion instead of a model representation that implicitly corresponds to that criterion.

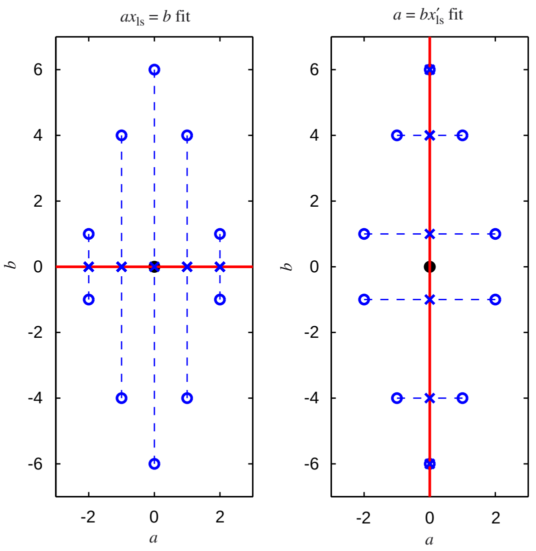

Fig. 1. The data and the least squares fits, minimizing vertical (left) and horizontal (right) distances.

TLS is an alternative to least squares for approximately solving an overdetermined system of equations. In terms of data fitting, the TLS method minimizes the sum of the squared orthogonal distances from the data points to the fitting line. Using the system of equations (1), line fitting by the TLS method leads to the problem

$$\min_{\substack{x\in\mathbb{R}\\ \widehat{a}_1,\ldots,\widehat{a}_N\in\mathbb{R}\\ \widehat{b}_1,\ldots,\widehat{b}_N\in\mathbb{R}}} \sum_{i=1}^{N} \left\| d_i - \begin{bmatrix} \widehat{a}_i \\ \widehat{b}_i \end{bmatrix} \right\|_2^2 \quad \text{s.t.} \quad \widehat{a}_i x = \widehat{b}_i \quad \text{for } i = 1, \ldots, N. \tag{5}$$

However, for the given data it has no solution. Informally, the TLS solution is $x_{\text{tls}} = \infty$, which corresponds to a fit by a vertical line. However, formally

> the TLS problem (5) has no solution for the data in the example and therefore does not give a fitting line.

By using (1) to define the TLS line fitting problem, we restrict the fitting line to be a graph of a function $ax = b$ for some $x \in \mathbb{R}$. Thus, we a priori exclude the vertical line as a possible solution. In the example, the line minimizing the sum of the squared orthogonal distances happens to be the vertical line. For this reason, x_{tls} does not exist.

Any line $\mathscr{B}$ passing through the origin can be represented as an image and a kernel, i.e., there exist matrices $P \in \mathbb{R}^{2\times 1}$ and $R \in \mathbb{R}^{1\times 2}$, such that

$$\mathscr{B} = \operatorname{image}(P) := \{d = Pl \in \mathbb{R}^2 \mid l \in \mathbb{R}\}$$

and

$$\mathscr{B} = \ker(R) := \{d \in \mathbb{R}^2 \mid Rd = 0\}.$$

Using an image representation, the problem of minimizing the sum of the orthogonal distances is

$$\min_{\substack{P\in\mathbb{R}^{2\times 1}\\ l_1,\ldots,l_N\in\mathbb{R}}} \sum_{i=1}^{N} \|d_i - \widehat{d}_i\|_2^2 \quad \text{s.t.} \quad \widehat{d}_i = Pl_i \quad \text{for } i = 1, \ldots, N. \tag{6}$$

With $D := [d_1 \ \cdots \ d_N]$, $\widehat{D} := [\widehat{d}_1 \ \cdots \ \widehat{d}_N]$, and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm, (6) is more compactly written as

$$\min_{\substack{P\in\mathbb{R}^{2\times 1}\\ L\in\mathbb{R}^{1\times N}}} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \widehat{D}=PL. \tag{7}$$

Similarly, using a kernel representation, we have

$$\min_{\substack{R\in\mathbb{R}^{1\times 2},R\neq 0\\ \widehat{D}\in\mathbb{R}^{2\times N}}} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad R\widehat{D}=0. \tag{8}$$

Contrary to the TLS problem (5), problems (7) and (8) always have (possibly nonunique) solutions. In the example, solutions are, e.g., $P^* = \mathrm{col}(0, 1)$ and $R^* = [1 \ 0]$, which describe the vertical line $\mathscr{B}^* := \mathrm{image}(P^*) = \ker(R^*)$.

The constraints $\widehat{D}=PL$, $P \in \mathbb{R}^{2\times 1}$, $L \in \mathbb{R}^{1\times N}$, and $R\widehat{D}=0$, $R \in \mathbb{R}^{1\times 2}$, $R \neq 0$ are equivalent to the constraint $\mathrm{rank}(\widehat{D}) \leqslant 1$, which shows that the points

> $\{\widehat{d}_1, \ldots, \widehat{d}_N\}$ being fitted exactly by a line passing through the origin is equivalent to $\mathrm{rank}([\widehat{d}_1 \ \cdots \ \widehat{d}_N]) \leqslant 1$.

In fact, (7) and (8) are instances of one and the same abstract problem: approximate the data matrix D by a rank-one matrix $\widehat{D}$.

1.2. Input/output interpretation of $AX = B$

The underlying goal is: given a set of points in $\mathbb{R}^{\mathrm{d}}$, find a subspace of $\mathbb{R}^{\mathrm{d}}$ of bounded dimension that has the least 2-norm distance to all points. Such a subspace is an optimal (in the 2-norm sense) fitting *model*. The most general way to *represent* any subspace in $\mathbb{R}^{\mathrm{d}}$ is the kernel or image formulation; the classical least squares and TLS formulations exclude some subspaces. As illustrated by the example, the equations $ax = b$ and $a = bx$, used in the least squares and TLS problem formulations to represent the subspace, might fail to represent the optimal solution, while the kernel and image representations do not have such deficiency. This suggests that the kernel and image representations are better suited for data fitting.

The equations $ax=b$ and $a=bx$ were introduced from an algorithmic point of view—by using them, the data fitting problem is turned into the standard problem of solving approximately an overdetermined system of equations. There is another, more insightful, interpretation of these equations that comes from system theory. In the model represented by the equation $ax=b$, the variable a is an input, meaning that it is free, and the variable b is an output, meaning that it is bound by the input and the model. Similarly, in the model represented by the equation $a = bx$, the variable a is an output and the variable b is an input. The input/output interpretation has an intuitive appeal because it shows a causal dependence of the variables: the input is causing the output.

Representing the model by an equation $ax = b$ or $a = bx$, as done in the classical method, one a priori assumes that the optimal fitting model has a certain input/output structure. The consequences are:

- existence of exceptional (nongeneric) cases, which complicate the theory,
- ill-conditioning caused by “nearly” exceptional cases, which leads to lack of numerical robustness of the algorithms, and
- need of regularization, which leads to a change of the specified fitting criterion.

These aspects of the classical method are generally considered as inherent to the data fitting problem. In fact, by choosing the alternative image and kernel model representations the nongeneric problems (and the related issues of ill-conditioning and need of regularization) are avoided.

1.3. Contributions of the paper and related work

Connections between data modeling problems and low-rank approximation—the topic of this paper—abound in the literature; however, often they are implicit and not deemed essential. The following are examples where the fact that an exact (noise-free) data matrix is low-rank is a common knowledge and is exploited in solution methods:

- realization theory—a sequence

  $$H = (H(0), H(1), \ldots, H(t), \ldots)$$

  is an impulse response of a linear time-invariant (LTI) system of order $\mathtt{n}$ if and only if the (infinite) Hankel matrix

  $$\mathscr{H}(H) := \begin{bmatrix} H(1) & H(2) & H(3) & \cdots \\ H(2) & H(3) & ⋰ & \\ H(3) & ⋰ & & \\ \vdots & & & \end{bmatrix}$$

  constructed from H has rank $\mathtt{n}$;
- direction-of-arrival problem in signal processing—the rank of an exact data matrix equals the number of sources;
- chemometrics—the rank of an exact data matrix equals the number of chemical components.

Although omnipresent, however, until now the structured low-rank approximation (SLRA) problem has not been generally perceived as a data modeling principle. It is the belief of the author that

> behind every linear data modeling problem there is a (hidden) low-rank approximation problem: the model imposes relations on the data which render a matrix constructed from exact data rank deficient.

Low-rank approximation is used in a number of data modeling problems from diverse scientific fields; however, there are problems, e.g., frequency domain and stochastic system identification, that are still awaiting for such an interpretation.

De Moor (1993, 1994), defines a generic problem, called structured TLS, and shows a number of applications that reduce to it. The structured TLS problem corresponds to the SLRA problem, defined in this paper (see Section 2.2) when a kernel representation is used to represent the rank constraint. We consider the more general SLRA problem because it gives freedom in choosing different representations for solving particular problems.

The rank constraint in the low-rank approximation problem corresponds to the constraint that the data are fitted by a linear model of bounded complexity in the data fitting problem. Therefore, the question of representing the rank constraint in the low-rank approximation problem corresponds to the question of choosing the model representation in the data fitting problem. The behavioral approach to system theory put forward by Willems (1986, 1987) is a manifestation of the representation free thinking. Deriving dynamic models from data, i.e., system identification, has been considered in the behavioral setting in Roorda and Heij (1995), Roorda (1995), and Markovsky, Willems, Van Huffel, De Moor, and Pintelon (2005).

The contributions of this paper are:

1. embed the structured TLS problem in the behavioral setting,
2. complete the list of applications in De Moor (1993) with output error identification, pole placement, harmonic retrieval, and approximate common divisor problems,
3. present generalizations and connections of the structured TLS problem to problems with nonnegativity constraint, rank minimization, and structured pseudospectra, and
4. give a tutorial to a representative set of data modeling problems from a unifying viewpoint.

1.4. Outline of the paper

Section 2 defines the low-rank approximation problem as a representation free data modeling problem, applying to general multivariable static and dynamic problems. Approximation by an *unstructured* matrix corresponds to fitting the data by a *static* linear model. Approximation by a Hankel *structured* matrix corresponds to fitting the data by a *dynamic* LTI model. The SLRA problem is further motivated in Section 3 by a list of applications from three major areas: system theory, signal processing, and computer algebra.

Algorithms for solving the SLRA problem are outlined in Section 4. First we state a well-known result that links a basic low-rank approximation problem—approximate the given matrix by an unstructured low-rank matrix in the Frobenius norm sense—to the singular value decomposition (SVD) of the data matrix. The general case can be approached using relaxations, that yield suboptimal solutions, local, or global optimization methods. We review the algorithms based on the variable projections and alternating least squares methods. In all approaches the structure in the data matrix can be exploited for achieving efficient computational methods.

Section 5 discusses other (apart from the structure preserving) generalizations of the basic low-rank approximation problem. They are classified under generalization of the cost function and additional constrains on the approximating matrix. Relation of the low-rank approximation problems to the rank minimization problem (RMP) and to the structured pseudospectra is explained.

2. Structured low-rank approximation as a data modeling problem

In Section 1.1, we illustrated the equivalence between line fitting and rank-one matrix approximation. In this section, we extend this equivalence to general linear static and dynamic data modeling problem. In the general case, the equivalent problem is SLRA.

Contrary to the common perception that a model is an equation, e.g., $AX=B$, we view a model as a set, e.g., a line passing through the origin in the line fitting problem. Appendix A collects basic facts, used in the paper, about LTI models and their representations, see also Polderman and Willems (1998) and Markovsky, Willems, Van Huffel, and De Moor (2006).

2.1. Unstructured low-rank approximation

The unstructured low-rank approximation problem is defined as follows.

Problem 1 (*Unstructured low-rank approximation*). *Given a matrix* $D \in \mathbb{R}^{\mathtt{d}\times N}$, *with* $\mathtt{d} \leqslant N$, *a matrix norm* $\|\cdot\|$, *and an integer* $\mathtt{m}$, $0<\mathtt{m}<\mathtt{d}$, *find a matrix*

$$\widehat{D}^* := \arg\min_{\widehat{D}} \quad \|D-\widehat{D}\| \quad \text{s.t.} \quad \operatorname{rank}(\widehat{D}) \leqslant \mathtt{m}.$$

The matrix $\widehat{D}^*$ is an optimal rank-$\mathtt{m}$ (or less) approximation of D with respect to the given norm $\|\cdot\|$.

A well-known early result on low-rank approximation is the Eckart–Young–Mirsky theorem (Eckart & Young, 1936). It gives a solution to the basic low-rank approximation problem (i.e., unstructured low-rank approximation problem with Frobenius norm) in terms of the SVD. This is a special case of Problem 1, when the norm $\|\cdot\|$ is the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. The Eckart–Young–Mirsky theorem and the closely related generic TLS algorithm are reviewed in Section 4.1.

The approximation $\widehat{D}$ being low-rank is equivalent to $\widehat{D}$ being generated by a linear model, so low-rank approximation can be given the interpretation of a data modeling problem. To show this, note that $\mathtt{m} := \operatorname{rank}(\widehat{D})$ being strictly less than the row dimension $\mathtt{d}$ of D is equivalent to the existence of a full rank matrix R with $\mathtt{p} := \mathtt{d}-\mathtt{m}$ rows, such that $R\widehat{D}=0$. Therefore, the columns $\widehat{d}_1,\ldots,\widehat{d}_N$ of $\widehat{D}$ obey $\mathtt{p}$ independent linear relations $r_j\widehat{d}_i=0$, given by the rows $r_1,\ldots,r_{\mathtt{p}}$ of R. The equation $R\widehat{d}=0$ is a kernel representation of the fitting model $\mathscr{B}$—an $\mathtt{m}$-dimensional subspace of the data space $\mathbb{R}^{\mathtt{d}}$.

The dimension $\mathtt{m}$ of the subspace $\mathscr{B} \subset \mathbb{R}^{\mathtt{d}}$ is a measure for the complexity of the model $\mathscr{B}$: the larger the subspace is, the more complicated (and therefore less useful) the model is.

However, the larger the subspace is, the better the fitting accuracy could be, so that there is a trade-off between complexity and accuracy. The data modeling problem that corresponds to Problem 1 bounds the complexity and maximizes the accuracy.

Problem 2 (*Static data modeling*). *Given N, $\mathtt{d}$-variable observations $\{d_1, \dots, d_N\} \subset \mathbb{R}^{\mathtt{d}}$, a matrix norm $\|\cdot\|$, and model complexity $\mathtt{m}$, $0 < \mathtt{m} < \mathtt{d}$, find an optimal approximate model*

$$\widehat{\mathscr{B}}^* := \arg \min_{\widehat{\mathscr{B}}, \widehat{D}} \quad \|D - \widehat{D}\|$$
$$\text{s.t.} \quad \operatorname{image}(\widehat{D}) \subseteq \widehat{\mathscr{B}} \text{ and } \dim(\widehat{\mathscr{B}}) \leqslant \mathtt{m}, \qquad (9)$$

where $D \in \mathbb{R}^{\mathtt{d} \times N}$ is the data matrix $D := [d_1 \ \cdots \ d_N]$.

The solution $\widehat{\mathscr{B}}^*$ is an optimal approximate model for the data D with complexity bounded by $\mathtt{m}$. Of course, $\widehat{\mathscr{B}}^*$ depends on the approximation criterion, specified by the given norm $\|\cdot\|$. A justification for the choice of the norm $\|\cdot\|$ is provided in the errors-in-variables setting.

In the errors-in-variables setting the data matrix D is assumed to be a noisy measurement of a true matrix $\bar{D}$

$$D = \bar{D} + \widetilde{D}, \quad \operatorname{image}(\bar{D}) = \bar{\mathscr{B}}, \quad \dim(\bar{\mathscr{B}}) \leqslant \mathtt{m},$$
$$\text{and } \operatorname{vec}(\widetilde{D}) \sim N(0, vW) \quad \text{where } W \succ 0, \ v > 0. \qquad (10)$$

(The notation "$W \succ 0$" is used for "W positive definite".) Here $\widetilde{D}$ is the measurement error that is assumed to be a random matrix with zero mean and normal distribution, and "vec" is the vectorization operator

$$\operatorname{vec}([\widetilde{d}_1 \ \cdots \ \widetilde{d}_N]) := \operatorname{col}(\widetilde{d}_1, \dots, \widetilde{d}_N).$$

The true matrix $\bar{D}$ is "generated" by a true model $\bar{\mathscr{B}} := \operatorname{image}(\bar{D})$, with a known complexity bound $\mathtt{m}$, which is the object to be estimated in the errors-in-variables setting.

Proposition 3 (*Maximum likelihood property of an optimal static model $\widehat{\mathscr{B}}^*$*). *Assume that the data are generated in the errors-in-variables setting* (10), *where the matrix $W \succ 0$ is known and the scalar v is unknown. Then a measurable solution $\widehat{\mathscr{B}}^*$ to Problem 2 with weighted 2-norm*

$$\|E\|_W := \sqrt{\operatorname{vec}^\top(E) W^{-1} \operatorname{vec}(E)} \quad \text{for all } E \qquad (11)$$

is a maximum likelihood estimator for the true model $\bar{\mathscr{B}}$.

The main assumption of Proposition 3 is $\operatorname{cov}(\operatorname{vec}(\widetilde{D})) = vW$, with W given. Note, however, that v is not given, so that the probability density function of $\widetilde{D}$ is not completely specified. Proposition 3 shows that the problem of computing the maximum likelihood estimator in the errors-in-variables setting is equivalent to Problem 1 with the weighted norm $\|\cdot\|_W$. This problem is called weighted low-rank approximation (WLRA) and is further considered in Section 5.1. In the special case, $W = I$, i.e., assuming that all entries of $\widetilde{D}$ are uncorrelated and identically distributed, the maximum likelihood estimator is given by the solution to the basic low-rank approximation problem. Maximum likelihood estimation for density functions other than normal leads to low-rank approximation with norms other than the weighted 2-norm; see Boyd and Vandenberghe (2004, Section 7.1.1) for the classical regression problem.

2.2. Structured low-rank approximation

SLRA is a low-rank approximation, in which the approximating matrix $\widehat{D}$ is required to have the same structure as the data matrix D. Typical structures encountered in applications are Hankel, Toeplitz, Sylvester, and circulant. In order to state the problem in its full generality, we first define a structured matrix. Consider a mapping $\mathscr{S}$ from a parameter space $\mathbb{R}^{n_p}$ to a set of matrices $\mathbb{R}^{m \times n}$. A matrix $\widehat{D} \in \mathbb{R}^{m \times n}$ is called *$\mathscr{S}$-structured* if it is in the image of $\mathscr{S}$, i.e., if there exists a parameter $\widehat{p} \in \mathbb{R}^{n_p}$, such that $\widehat{D} = \mathscr{S}(\widehat{p})$.

Remark 4 (*Nonlinearly structured matrices*). Nonlinearly structured matrices are not considered in this paper because the corresponding nonlinearly SLRA problems are much harder to solve than the affine ones and in the errors-in-variables setting the corresponding maximum likelihood estimators are inconsistent.

SLRA Problem. *Given a structure specification $\mathscr{S} : \mathbb{R}^{n_p} \to \mathbb{R}^{m \times n}$, with $m \leqslant n$, a parameter vector $p \in \mathbb{R}^{n_p}$, a vector norm $\|\cdot\|$, and an integer r, $0 < r < \min(m, n)$, find a vector*

$$\widehat{p}^* := \arg \min_{\widehat{p}} \quad \|p - \widehat{p}\|$$
$$\text{s.t.} \quad \operatorname{rank}(\mathscr{S}(\widehat{p})) \leqslant r. \qquad (12)$$

The matrix $\widehat{D}^* := \mathscr{S}(\widehat{p}^*)$ is an optimal rank-r (or less) approximation of $D := \mathscr{S}(p)$, within the class of matrices with the same structure as D. Obviously, Problem 1 is a special case of the SLRA problem.

The reason to consider the more general structured low-rank approximation is that $D = \mathscr{S}(p)$ being low-rank and Hankel structured is equivalent to p being generated by an LTI dynamic model. To show this, consider first the special case of a scalar Hankel structure

$$\mathscr{H}_{\ell+1}(p) := \begin{bmatrix} p_1 & p_2 & \cdots & p_{n_p-\ell} \\ p_2 & p_3 & \cdots & p_{n_p-\ell+1} \\ \vdots & \vdots & & \vdots \\ p_{\ell+1} & p_{\ell+2} & \cdots & p_{n_p} \end{bmatrix}.$$

The approximation matrix $\widehat{D} = \mathscr{H}_{\ell+1}(\widehat{p})$ being rank deficient implies that there is a nonzero vector $R = [R_0 \ R_1 \ \cdots \ R_\ell]$, such that $R\mathscr{H}_{\ell+1}(\widehat{p}) = 0$. Due to the Hankel structure, this equation can be written as

$$R_0 \widehat{p}_t + R_1 \widehat{p}_{t+1} + \cdots + R_\ell \widehat{p}_{t+\ell} = 0 \quad \text{for } t = 1, \dots, n_p - \ell.$$

The homogeneous constant coefficients difference equation

$$R_0 w(t) + R_1 w(t+1) + \cdots + R_\ell w(t+\ell) = 0$$
$$\text{for } t = 1, 2, \dots \qquad (13)$$

describes an autonomous LTI system $\mathscr{B}$. More precisely, $\mathscr{B}$ is the solution set to (13), i.e.,

$$\mathscr{B} = \mathscr{B}(R) := \{w \in \mathbb{R}^{\mathbb{N}} \mid (13) \text{ holds}\}.$$

Let $\mathscr{B}_{\overline{[1,T]}}$ be the restriction of $\mathscr{B}$ on the interval $[1, \ldots, T]$, i.e.,

$$\mathscr{B}_{\overline{[1,T]}} := \{w \in \mathbb{R}^{T} \mid \text{there exists } w_{\mathrm{f}}, \text{ such that } (w, w_{\mathrm{f}}) \in \mathscr{B}\},$$

and note that for an autonomous system $\mathscr{B}$, $\dim(\mathscr{B}_{\overline{[1,T]}}) = \ell$, for all $T \geqslant \ell$, where ℓ is the lag of the difference equation (13). As in the static case, $\dim(\mathscr{B})$ is a measure for the complexity of the model.

The scalar Hankel low-rank approximation problem is then equivalent to the following signal modeling problem. Given T samples of scalar signal $w_{\mathrm{d}} \in \mathbb{R}^{T}$ (the subscript d stands for "data"), a signal norm $\|\cdot\|$, and a model complexity ℓ, find an optimal approximate model

$$\widehat{\mathscr{B}}^{*} := \arg\min_{\widehat{\mathscr{B}}, \widehat{w}} \quad \|w_{\mathrm{d}} - \widehat{w}\| \qquad \text{s.t.} \quad \widehat{w} \in \widehat{\mathscr{B}}_{\overline{[1,T]}} \text{ and } \dim(\widehat{\mathscr{B}}_{\overline{[1,T]}}) \leqslant \ell. \tag{14}$$

The solution $\widehat{\mathscr{B}}^{*}$ is an optimal approximate model for the signal w_{d} with bounded complexity: lag at most ℓ.

In the general case when the signal w is vector valued with $\mathtt{w}$ variables, the model $\mathscr{B}$ can be represented by a difference equation (13), where the parameters R_i are $g \times \mathtt{w}$ matrices. It turns out that for full rank polynomial matrix $R(z) := \sum_{i=0}^{\ell} z^{i} R_{i}$, the row dimension g of R is equal to the number of outputs $\mathtt{p}$ of the model (Willems, 1991, Proposition VIII.6). Correspondingly $\mathtt{m} := \mathtt{w} - \mathtt{p}$ is the number of inputs. For a general LTI system $\mathscr{B}$

$$\dim(\mathscr{B}_{\overline{[1,T]}}) \leqslant \mathtt{m}T + \ell\mathtt{p} \quad \text{for } T \geqslant \ell. \tag{15}$$

Thus the complexity of a general LTI model is specified by the pair of integers $(\mathtt{m}, \ell)$. Let $\mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell}$ be the class of bounded complexity LTI systems with $\mathtt{w}$ external variables, at most $\mathtt{m}$ inputs, and lag at most ℓ. The block-Hankel structured low-rank approximation problem is equivalent to the following LTI dynamic modeling problem.

Problem 5 (*LTI dynamic modeling problem*). *Given T samples,* $\mathtt{w}$ *variables, vector signal* $w_{\mathrm{d}} \in (\mathbb{R}^{\mathtt{w}})^{\mathrm{T}}$, *a signal norm* $\|\cdot\|$, *and a model complexity* $(\mathtt{m}, \ell)$, *find an optimal approximate model*

$$\widehat{\mathscr{B}}^{*} := \arg\min_{\widehat{\mathscr{B}}, \widehat{w}} \quad \|w_{\mathrm{d}} - \widehat{w}\| \qquad s.t. \quad \widehat{w} \in \widehat{\mathscr{B}}_{\overline{[1,T]}} \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell}. \tag{16}$$

The solution $\widehat{\mathscr{B}}^{*}$ is an optimal approximate model for the signal w_{d} with complexity bounded by $(\mathtt{m}, \ell)$. Note that (16) reduces to (14) when $\mathtt{m}=0$, i.e., when the model is autonomous, and to (9) when $\ell = 0$, i.e., when the model is static.

Similar to the static modeling problem, the dynamic modeling problem has a maximum likelihood interpretation in the errors-in-variables setting.

Proposition 6 (*Maximum likelihood property of an optimal dynamic model* $\widehat{\mathscr{B}}^{*}$). *Assume that the data* w_{d} *are generated in the errors-in-variables setting*

$$w_{\mathrm{d}} = \bar{w} + \widetilde{w} \quad \text{where } \bar{w} \in \bar{\mathscr{B}}_{\overline{[1,T]}} \in \mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell} \text{ and } \widetilde{w} \sim N(0, vI).$$

Then an optimal approximate model $\widehat{\mathscr{B}}^{*}$, *solving* (16) *with* $\|\cdot\| = \|\cdot\|_{2}$ *is a maximum likelihood estimator for the true model* $\bar{\mathscr{B}}$.

Except for a few special cases, see Section 4.1, currently there is no method that solves the SLRA problem globally and efficiently. In Section 4, we present local optimization methods and describe how the structure in the data matrix can be exploited for efficient cost function evaluation.

3. Applications

In this section we show applications of SLRA in system theory, signal processing, and computer algebra. Different applications lead to different types of structure $\mathscr{S}$. In most applications, however, $\mathscr{S}$ is composed of one or two blocks that are Hankel, unstructured, or fixed. (A block being fixed means that it is not modified in the search for the optimal approximation. A problem with Toeplitz structured blocks can be reformulated as an equivalent problem with Hankel structured blocks by rearranging the rows of the data matrix.) Consequently, algorithms and software for solving SLRA problems with such flexible structure specification can be readily used in the applications.

The presented applications are:

- System and control theory:
  1. Errors-in-variables system identification.
  2. Approximate realization.
  3. Model reduction.
  4. Output error system identification.
  5. Low-order controller design.
- Signal processing:
  6. Output only system identification.
  7. Finite impulse response (FIR) system identification.
  8 Harmonic retrieval.
- Computer algebra:
  9. Approximate greatest common divisor.

Most of the work on the errors-in-variables identification problem (see Söderström, 2007 and the references there in) is presented in the classical input/output setting, i.e., the proposed methods aim to derive a transfer function, matrix fraction description, or input/state/output representation of the system. The salient feature of the errors-in-variables problems, however, is that all variables are treated on an equal footing as noise corrupted. Therefore, the input/output partitioning implied by the classical model representations is irrelevant in this problem. Section 3.1 relates the LTI dynamic modeling problem 5 to the errors-in-variables identification problem, posed in a representation free setting.

Contrary to the errors-in-variables problem, the applications presented in Sections 3.2–3.5 and 3.7 do assume a given input/output partitioning. The approximate realization (Section 3.2) and model reduction (Section 3.3) problems approximate, respectively, a given noisy impulse response and an impulse response of a high order LTI system (and of course the impulse response depends on a specified input/output partition). The output error identification problem (Section 3.4) imposes the constraint that part of the variables are noise free and the controller design problem (Section 3.5) involves a feedback interconnection, which also assume a given input/output structure of the model.

3.1. Errors-in-variables identification

Proposition 6 shows that the maximum likelihood estimate $\widehat{\mathscr{B}}^*$ of the true model $\bar{\mathscr{B}}$ in the errors-in-variables setting is defined by a SLRA problem with Hankel SLRA problem with Hankel structured data matrix $\mathscr{S}(p) = \mathscr{H}_{\ell+1}(w_\mathrm{d})$ and rank reduction with the number of outputs $\mathtt{p}$. Under additional stochastic assumptions, see Pintelon and Schoukens (2001), Kukush, Markovsky, and Van Huffel (2005), the estimator $\widehat{\mathscr{B}}^*$ is consistent and the estimated parameters have asymptotically normal joint distribution. This allows us to compare asymptotic confidence regions, i.e., the probability that the true parameters lie inside the confidence region tends to a prescribed value, as the sample size tends to infinity.

The statistical setting gives a recipe for choosing the norm $\|\cdot\|$ and a "quality certificate" for the approximation method (16): the method works "well" (consistency) and is optimal (asymptotic efficiency[1]) under certain specified conditions. However, the assumption that the data are generated by a true model with additive noise is sometimes not realistic. Model-data mismatch is often due to a restrictive LTI model class being used and not (only) due to measurement noise. This implies that the approximation aspect of the method is often more important than the stochastic estimation one.

The following problems can also be given the interpretation of defining maximum likelihood estimators under appropriate stochastic assumptions. However, we do not do this and give only their deterministic definitions.

3.2. Approximate realization

Define the 2-norm $\|\Delta H\|_2$ of a matrix-valued signal $\Delta H \in (\mathbb{R}^{\mathtt{p}\times\mathtt{m}})^{T+1}$ as $\|\Delta H\|_2 := \sqrt{\sum_{t=0}^{T}\|\Delta H(t)\|_\mathrm{F}^2}$, and let σ be the shift operator $\sigma(H)(t) = H(t+1)$. Acting on a finite time series $(H(0), H(1), \ldots, H(T))$, σ deletes the first sample $H(0)$.

Problem 7 (*Approximate realization*). *Given $H_\mathrm{d} \in (\mathbb{R}^{\mathtt{p}\times\mathtt{m}})^T$ and a complexity specification ℓ, find an optimal approximate model for H_d of a bounded complexity $(\mathtt{m}, \ell)$*

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{H},\widehat{\mathscr{B}}} \quad \|H_\mathrm{d} - \widehat{H}\|_2 \quad \text{s.t.} \quad \widehat{H} \text{ is the impulse response of } \widehat{\mathscr{B}} \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{m}+\mathtt{p}}_{\mathtt{m},\ell}.$$

Proposition 8. *Problem 7 is equivalent to the SLRA problem, with $\|\cdot\| = \|\cdot\|_2$, Hankel structured data matrix $\mathscr{S}(p) = \mathscr{H}_{\ell+1}(\sigma H_\mathrm{d})$, and rank reduction by the number of outputs $\mathtt{p}$.*

Approximate realization is a special identification problem. The input is a pulse and the initial conditions are zeros. Nevertheless, the exact version of this problem is a very much studied problem. The classical references are the Ho and Kalman's (1966) realization algorithm and Kung's (1978) algorithm.

It can be shown that the optimal approximate model $\widehat{\mathscr{B}}^*$ does not depend on the shape of the Hankel matrix as long as the Hankel matrix dimensions are sufficiently large (at least $\mathtt{p}(\ell+1)$ rows and at least $\mathtt{m}(\ell+1)$ columns). However, solving the low-rank approximation problem for a data matrix $\mathscr{H}_{L+1}(\sigma H_\mathrm{d})$, where $L > \ell$, one needs to achieve rank reduction by $\mathtt{p}(L-\ell+1)$ instead of by $\mathtt{p}$ as in Proposition 8. Larger rank reduction leads to more difficult computational problems. On one hand, the cost per iteration gets higher and on another hand, the search space gets higher dimensional, which makes the optimization algorithm more susceptible to local minima.

3.3. Model reduction

The finite time-T H_2 norm $\|\Delta\mathscr{B}\|_{2,T}$ of an LTI system $\Delta\mathscr{B}$ is defined as the 2-norm of the sequence of its first T Markov parameters, i.e., if ΔH is the impulse response of $\Delta\mathscr{B}$, $\|\Delta\mathscr{B}\|_{2,T} := \|\Delta H\|_2$.

Problem 9 (*Finite time* H_2 *model reduction*). *Given an LTI system $\mathscr{B}_\mathrm{d} \in \mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell}$ and a complexity specification $\ell_\mathrm{red} < \ell$, find an optimal approximation of $\mathscr{B}_\mathrm{d}$ with bounded complexity $(\mathtt{m}, \ell_\mathrm{red})$.*

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}} \quad \|\mathscr{B}_\mathrm{d} - \widehat{\mathscr{B}}\|_{2,T} \quad \text{s.t.} \quad \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell_\mathrm{red}}.$$

Proposition 10. *Problem 9 is equivalent to the SLRA problem with $\|\cdot\| = \|\cdot\|_2$, Hankel structured data matrix $\mathscr{S}(p) = \mathscr{H}_{\ell+1}(H_\mathrm{d})$, where H_d is the impulse response of $\mathscr{B}_\mathrm{d}$, and rank reduction by the number of outputs $\mathtt{p} := \mathtt{w} - \mathtt{m}$.*

Finite time H_2 model reduction is equivalent to the approximate realization problem with H_d being the impulse response of $\mathscr{B}_\mathrm{d}$. In practice, $\mathscr{B}_\mathrm{d}$ need not be linear since in the model reduction problem only the knowledge of its impulse response H_d is used. If H_d is LTI and stable, the approximation $\widehat{\mathscr{B}}^*$

[1] In the errors-in-variables setting the maximum likelihood estimator does not have an expected value, however, for linear models it has the smallest possible asymptotic covariance matrix.

converges, as $T \to \infty$, to the optimal approximation in the 2-norm sense.

3.4. Output error identification

Problem 11 (*Output error identification*). *Given a signal $w_{\mathrm{d}} := (u_{\mathrm{d}}, y_{\mathrm{d}}) \in (\mathbb{R}^{\mathtt{m}} \times \mathbb{R}^{\mathtt{p}})^T$ with an input/output partitioning and a complexity specification ℓ, find an optimal approximate model for w_{d} of a bounded complexity $(\mathtt{m}, \ell)$*

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}, \widehat{y}} \quad \|y_{\mathrm{d}} - \widehat{y}\|_2$$

$$s.t. \quad (u_{\mathrm{d}}, \widehat{y}) \in \widehat{\mathscr{B}}|_{\overline{[1,T]}} \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{m}+\mathtt{p}}_{\mathtt{m},\ell}.$$

Proposition 12. *Problem 11 is equivalent to the SLRA problem with $\|\cdot\| = \|\cdot\|_2$, data matrix*

$$\mathscr{S}(p) = \begin{bmatrix} \mathscr{H}_{\ell+1}(u_{\mathrm{d}}) \\ \mathscr{H}_{\ell+1}(y_{\mathrm{d}}) \end{bmatrix}$$

composed of a fixed block and a Hankel structured block, and rank reduction by the number of outputs $\mathtt{p}$.

Output error identification is one of the standard system identification problems (Ljung, 1999; Söderström & Stoica, 1989). It is a special case of the prediction error methods when the noise term is not modeled.

3.5. Pole placement by a low-order controller

Consider the feedback system shown in Fig. 2. For simplicity we restrict to the single input, single output case. The polynomials P and Q, define the transfer function Q/P of the plant and are given. They are assumed to be relatively prime and the transfer function Q/P is assumed to satisfy the constraint $\deg(Q) \leqslant \deg(P) =: \ell_P$, which ensures that the plant is a causal LTI system. The polynomials Y and X describe the controller and are unknowns. The design constraints are that the controller should be causal and have order bounded by a specified integer ℓ_P. These specifications translate to the following constraints on the polynomials Y and X

$$\deg(Y) \leqslant \deg(X) =: \ell_X < \ell_P. \tag{17}$$

The pole placement problem is to determine X and Y, so that the poles of the closed-loop system are as close as possible in some specified sense to desired locations, given by the roots of a polynomial F, where $\deg(F) = \ell_X + \ell_P$. We consider a problem that aims to assign exactly the poles of a plant that is as close as possible to the given plant.

In what follows, we use the correspondence between $\ell_P + 1$ dimensional vectors and ℓ_Pth degree polynomials

$$\operatorname{col}(P_0, P_1, \ldots, P_{\ell_P}) \in \mathbb{R}^{\ell_P+1}$$

$$\leftrightarrow \quad P(z) = P_0 + P_1 z + \cdots + P_{\ell_P} z^{\ell_P} \in \mathbb{R}[z] \tag{18}$$

and (with some abuse of notation) refer to P as both a vector and a polynomial.

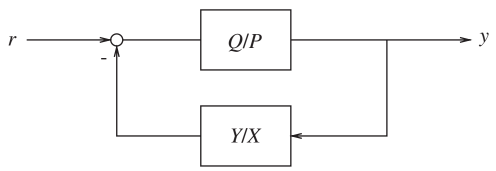

Fig. 2. Feedback control system.

Problem 13 (*Pole placement by low-order controller*). *Given*

1. *the transfer function Q/P of a plant,*
2. *a polynomial F, whose roots are the desired poles of the closed-loop system, and*
3. *a bound $\ell_X < \deg(P)$ on the order of the controller,*

find the transfer function Y/X of the controller, such that

1. *the degree constraint (17) is satisfied and*
2. *the controller assigns the poles of a system whose transfer function $\widehat{Q}/\widehat{P}$ is as close as possible to the transfer function Q/P in the sense that*

   $$\|\operatorname{col}(P, Q) - \operatorname{col}(\widehat{P}, \widehat{Q})\|_2$$

   is minimized.

Next, we write down explicitly the considered optimization problem, which shows its equivalence to a SLRA problem. The closed-loop transfer function is $QX/(PX + QY)$, so that a solution to the pole placement problem is given by a solution to the Diophantine equation

$$PX + QY = F. \tag{19}$$

Eq. (19) can be written as a Sylvester structured system of equations

$$\underbrace{\begin{bmatrix} P_0 & & & Q_0 & & \\ P_1 & \ddots & & Q_1 & \ddots & \\ \vdots & \ddots & P_0 & \vdots & \ddots & Q_0 \\ P_{\ell_P} & & P_1 & Q_{\ell_P} & & Q_1 \\ & \ddots & \vdots & & \ddots & \vdots \\ & & P_{\ell_P} & & & Q_{\ell_P} \end{bmatrix}}_{\mathscr{R}_{\ell_X+1}(P,Q)} \begin{bmatrix} X_0 \\ \vdots \\ X_{\ell_X} \\ Y_0 \\ \vdots \\ Y_{\ell_X} \end{bmatrix} = \underbrace{\begin{bmatrix} F_0 \\ \vdots \\ F_{\ell_P} \\ F_{\ell_P+1} \\ \vdots \\ F_{\ell_P+\ell_X} \end{bmatrix}}_{F},$$

which is an overdetermined system of equations due to the degree constraint (17). Therefore, problem 13 can be written as

$$\min_{\substack{\widehat{P},\widehat{Q} \in \mathbb{R}^{\ell_P+1} \\ X,Y \in \mathbb{R}^{\ell_X+1}}} \left\| \begin{bmatrix} P \\ Q \end{bmatrix} - \begin{bmatrix} \widehat{P} \\ \widehat{Q} \end{bmatrix} \right\|_2$$

$$\text{s.t.} \quad \mathscr{R}_{\ell_X+1}(\widehat{P}, \widehat{Q}) \begin{bmatrix} X \\ Y \end{bmatrix} = F.$$

Proposition 14. *Problem* 13 *is equivalent to the SLRA problem with* $\|\cdot\| = \|\cdot\|_2$, *data matrix*

$$\mathscr{S}(p) = \begin{bmatrix} [F_0 \;\; F_1 \;\; \cdots \;\; F_{\ell_P+\ell_X}] \\ \mathscr{R}^\top_{\ell_X+1}(P,Q) \end{bmatrix}$$

composed of a fixed block and a Sylvester structured block, *and rank reduction by* 1.

3.6. Output only identification

The model class of autonomous LTI systems is $\mathscr{L}^{\mathtt{p}}_{0,\ell}$. Excluding the cases of multiple poles, $\mathscr{L}^{\mathtt{p}}_{0,\ell}$ is equivalent to the *sum-of-damped exponentials model* class, i.e., signals y that can be represented in the form

$$y(t) = \sum_{j=1}^{\ell} a_j \mathrm{e}^{d_j t} \mathrm{e}^{\mathbf{i}(\omega_j t + \phi_j)} \quad (\mathbf{i} := \sqrt{-1}).$$

The parameters $\{a_j, d_j, \omega_j, \phi_j\}_{j=1}^{\ell}$ of the sum-of-damped exponentials model have the following meaning: a_j are amplitudes, d_j damping, ω_j frequencies, and ϕ_j initial phases.

Problem 15 (*Output only identification*). *Given a signal* $y_{\mathrm{d}} \in (\mathbb{R}^{\mathtt{p}})^T$ *and a complexity specification* ℓ, *find an optimal approximate model for* y_{d} *of bounded complexity* $(0, \ell)$

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}},\widehat{y}} \quad \|y_{\mathrm{d}} - \widehat{y}\|_2 \quad \text{s.t.} \quad \widehat{y} \in \widehat{\mathscr{B}}_{\overline{[1,T]}} \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{p}}_{0,\ell}.$$

Proposition 16. *Problem* 15 *is equivalent to the SLRA problem with* $\|\cdot\| = \|\cdot\|_2$, *a Hankel structured data matrix* $\mathscr{S}(p) = \mathscr{H}_{\ell+1}(y_{\mathrm{d}})$, *and rank reduction by the number of outputs* $\mathtt{p}$.

Output only identification is equivalent to approximate realization and finite time H_2 model reduction. In the signal processing literature, the problem is known as *linear prediction*.

3.7. Finite impulse response system identification

Denote by $\mathrm{FIR}_{\mathtt{m},\ell}$ the model class of finite impulse response LTI systems with at most $\mathtt{m}$ inputs and lag at most ℓ, i.e.,

$$\mathrm{FIR}_{\mathtt{m},\ell} := \{\mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell} \mid \mathscr{B} \text{ has finite impulse response}\}.$$

Problem 17 (*FIR identification*). *Given a signal* $w_{\mathrm{d}} := (u_{\mathrm{d}}, y_{\mathrm{d}}) \in (\mathbb{R}^{\mathtt{m}} \times \mathbb{R}^{\mathtt{p}})^T$ *with an input/output partition and a complexity specification* ℓ, *find an optimal approximate FIR model for* w_{d} *of bounded complexity* $(\mathtt{m}, \ell)$

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}},\widehat{w}} \quad \|w_{\mathrm{d}} - \widehat{w}\|_2 \quad \text{s.t.} \quad \widehat{w} \in \widehat{\mathscr{B}}_{\overline{[1,T]}} \text{ and } \widehat{\mathscr{B}} \in \mathrm{FIR}_{\mathtt{m},\ell}.$$

Proposition 18. *Problem* 17 *is equivalent to the SLRA problem with* $\|\cdot\| = \|\cdot\|_2$, *data matrix*

$$\mathscr{S}(p) = \begin{bmatrix} [y_{\mathrm{d}}(1) \;\; \cdots \;\; y_{\mathrm{d}}(T-\ell)] \\ \mathscr{H}_{\ell+1}(u_{\mathrm{d}}) \end{bmatrix},$$

composed of a fixed block and a Hankel structured block, *and rank reduction by the number of outputs* $\mathtt{p}$.

For exact data, i.e., assuming that

$$y_{\mathrm{d}}(t) = (h \star u_{\mathrm{d}})(t) := \sum_{\tau=0}^{\ell} h(\tau) u_{\mathrm{d}}(t-\tau)$$

the FIR identification problem is equivalent to the deconvolution problem: given the signals u_{d} and $y_{\mathrm{d}} := H \star u_{\mathrm{d}}$, find the signal H. For noisy data, the FIR identification problem can be viewed as an *approximate deconvolution problem*. The approximation is in the sense of finding the nearest signals $\widehat{u}$ and $\widehat{y}$ to the given ones u_{d} and y_{d}, such that $\widehat{y} := \widehat{H} \star \widehat{u}$, for a signal $\widehat{H}$ with a given length ℓ.

3.8. Harmonic retrieval

The aim of the harmonic retrieval problem is to approximate the data by a sum of sinusoids. From a system theoretic point of view, harmonic retrieval aims to approximate the data by a marginally stable autonomous model.

Problem 19 (*Harmonic retrieval*). *Given a signal* $y_{\mathrm{d}} \in (\mathbb{R}^{\mathtt{p}})^T$ *and a complexity specification* ℓ, *find an optimal approximate model for* y_{d} *that is in the model class* $\mathscr{L}^{\mathtt{p}}_{0,\ell}$ *and is marginally stable*

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}},\widehat{y}} \quad \|y_{\mathrm{d}} - \widehat{y}\|_2 \quad \text{s.t.} \quad \widehat{y} \in \widehat{\mathscr{B}}_{\overline{[1,T]}}, \quad \widehat{\mathscr{B}} \in \mathscr{L}^{\mathtt{p}}_{0,\ell}, \text{ and } \widehat{\mathscr{B}} \text{ is marginally stable.}$$

Due to the stability constraint, Problem (19) is not a special case of the SLRA problem. In the univariate case $\mathtt{p} = 1$, however, a necessary condition for an autonomous model $\mathscr{B}$ to be marginally stable is that a kernel representation $\mathscr{B}(R)$ of $\mathscr{B}$ is either palindromic,

$$R(z) := \sum_{i=0}^{\ell} z^i R_i \text{ is palindromic} \quad :\Longleftrightarrow \quad R_{\ell-i} = R_i \quad \text{for } i = 0, 1, \ldots, \ell$$

or antipalindromic: $R_{\ell-i} = -R_i$, for $i = 0, 1, \ldots, \ell$. The antipalindromic case is nongeneric in the space of the marginally stable systems, so as relaxation of the stability constraint, we can use the constraint that the kernel representation is palindromic.

Problem 20 (*Harmonic retrieval, relaxed version, and scalar case*). *Given a signal* $y_{\mathrm{d}} \in (\mathbb{R})^T$ *and a complexity* *specification ℓ, find an optimal approximate model for y_{d} that is in the model class $\mathscr{L}^1_{0,\ell}$ and has a palindromic kernel representation*

$$\begin{aligned} \widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}},\widehat{y}} \quad & \|y_{\mathrm{d}} - \widehat{y}\|_2 \\ s.t. \quad & \widehat{y} \in \widehat{\mathscr{B}}|_{[1,T]}, \\ & \widehat{\mathscr{B}} \in \mathscr{L}^1_{0,\ell} \textit{ and } \widehat{\mathscr{B}}(\widehat{R}) = \widehat{\mathscr{B}} \textit{ are palindromic.} \end{aligned}$$

The constraint that R is palindromic can be expressed as a structural constraint on the data matrix, which reduces the relaxed harmonic retrieval problem to the SLRA problem.

Proposition 21. *Problem* 20 *is equivalent to the SLRA problem with $\|\cdot\| = \|\cdot\|_2$, structured data matrix composed of a Hankel next to a Toeplitz block*

$$\mathscr{S}(p) = [\mathscr{H}_{\ell+1}(y) \quad \mathscr{T}_{\ell+1}(y)],$$

where

$$\mathscr{T}_{\ell+1}(y) := \begin{bmatrix} y_{\ell+1} & y_{\ell+2} & \cdots & y_T \\ \vdots & \vdots & & \vdots \\ y_2 & y_3 & \cdots & y_{T-\ell+1} \\ y_1 & y_2 & \cdots & y_{T-\ell} \end{bmatrix},$$

and rank reduction by 1.

3.9. Approximate common divisor

Let $\mathrm{GCD}(a,b)$ be the greatest common divisor of the polynomials a and b and recall the one-to-one correspondence (18) between vectors in $\mathbb{R}^{n+1}$ and nth degree polynomials.

Problem 22 (*Approximate common divisor*). *Given vectors $a, b \in \mathbb{R}^{n+1}$ and an integer $d \in \mathbb{N}$, find a vector*

$$\begin{aligned} \widehat{c}^* = \arg \min_{\substack{\widehat{a},\widehat{b}\in\mathbb{R}^{n+1} \\ \widehat{c}\in\mathbb{R}^{d+1}}} \quad & \|\mathrm{col}(a,b) - \mathrm{col}(\widehat{a},\widehat{b})\|_2 \\ s.t. \quad & \widehat{c} = GCD(\widehat{a},\widehat{b}) \textit{ and } \deg(\widehat{c}) = d. \end{aligned}$$

Proposition 23. *Problem* 22 *is equivalent to the SLRA problem with $\|\cdot\| = \|\cdot\|_2$, Sylvester structure $\mathscr{S}(p) = \mathscr{R}^\top_{n-d+1}(a,b)$, and rank reduction by* 1.

For $\mathtt{p}\times\mathtt{m}$ matrix polynomials, the structure is block-Sylvester and the necessary rank reduction is by $\mathtt{p}$. For two variable polynomial, the structure is block-Sylvester–Sylvester-block.

4. Algorithms

A few special SLRA problems have analytic solutions, see Section 4.1, however in general the SLRA problem is NP-hard. There are three fundamentally different approaches for solving it: convex relaxations, see Section 4.2, local optimization, see Section 4.3, and global optimization. The approach that is currently most developed (and that we describe in most details) is the one using local optimization methods. Section 4.4 shows a simulation example of data fitting using the SLRA paradigm.

4.1. Special cases with known analytic solutions

The Eckart–Young–Mirsky theorem gives a solution to the unstructured low-rank approximation problem with Frobenius norm criterion in terms of the SVD.

Theorem 24 (*Eckart–Young–Mirsky theorem*). *Let $D = U\Sigma V^\top$ be the SVD of $D \in \mathbb{R}^{\mathtt{d}\times N}$ and partition the matrices U, Σ, and V as follows*:

$$\mathrm{U} =: \begin{array}{c} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} \\ [\,U_1 \quad U_2,\,] \end{array}', \quad \Sigma =: \begin{array}{cc} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & \begin{array}{c} \mathtt{m} \\ \mathtt{p} \end{array} \end{array}, \quad \mathrm{V} =: \begin{array}{c} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} \\ [\,V_1 \quad V_2\,] \end{array}',$$

where $\mathtt{m} \in \mathbb{N}$, $0 \leqslant \mathtt{m} \leqslant \min(\mathtt{d}, N)$, and $\mathtt{p} := \mathtt{d} - \mathtt{m}$. Then the rank-$\mathtt{m}$ matrix

$$\widehat{D}^* = U_1\Sigma_1V_1^\top$$

is such that

$$\|D - \widehat{D}^*\|_{\mathrm{F}} = \min_{\mathrm{rank}(\widehat{D}) \leqslant \mathtt{m}} \|D - \widehat{D}\|_{\mathrm{F}} = \sqrt{\sigma^2_{\mathtt{m}+1} + \cdots + \sigma^2_{\mathtt{d}}},$$

where $\mathrm{diag}(\sigma_1, \ldots, \sigma_{\mathtt{d}}) := \Sigma$. The solution $\widehat{D}^*$ is unique if and only if $\sigma_{\mathtt{m}+1} \neq \sigma_{\mathtt{m}}$.

As shown in Vanluyten, Willems, and De Moor (2005), the solution D^* is optimal with respect to any norm $\|\cdot\|$ that is invariant under orthogonal transformations, i.e., satisfying the relation $\|UDV\| = \|D\|$, for any D and for any orthogonal matrices U and V. Moreover, Theorem 24 can be generalized to weighted norms of the form $\|W_{\mathrm{l}}(D - \widehat{D})W_{\mathrm{r}}\|$, where W_{l} and W_{r} are positive definite weight matrices, see Section 5.1. These are the most general unstructured WLRA problems that are known to have analytic solution in terms of the SVD.

Closely related to the basic low-rank approximation problem is the TLS problem: given matrices $A \in \mathbb{R}^{N\times\mathtt{m}}$ and $B \in \mathbb{R}^{N\times\mathtt{p}}$, solve the optimization problem

$$\begin{aligned} \min_{\widehat{A},\widehat{B},X} \quad & \|[A \;\; B] - [\widehat{A} \;\; \widehat{B}]\|_{\mathrm{F}} \\ \text{s.t.} \quad & \widehat{A}X = \widehat{B}. \end{aligned}$$

The TLS problem is put forward in Golub and Van Loan (1980) for the case when B is a vector (system of equations with one right-hand-side). The general case is treated in the monograph (Van Huffel & Vandewalle, 1991).

Theorem 25 (*Solution to the TLS problem*). *Let*

$$[A \;\; B] = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_{\mathtt{m}+\mathtt{p}})V^\top$$

be the SVD of $[A \;\; B]$ *and partition the matrix V as follows*

$$\mathrm{V}:=\begin{array}{cc} \mathtt{m} & \mathtt{p} \\ {[\,V_1} & V_2,\,] \end{array} =: \begin{array}{cc} \mathtt{m} & \mathtt{p} \\ \left[\begin{array}{cc} V_{11} & V_{12} \\ V_{21} & V_{22} \end{array}\right] \end{array} \begin{array}{c} \mathtt{m} \\ \mathtt{p} \end{array}.$$

A TLS solution exists if and only if the matrix V_{22} *is nonsingular. In this case, a solution is*

$$X_{\text{tls}} = -V_{12}V_{22}^{-1}.$$

It is unique if and only if $\sigma_{\mathtt{m}+1} \neq \sigma_{\mathtt{m}}$.

As the solution to the basic low-rank approximation problem, the solution to the TLS problem is also based on the SVD of the data matrix $D^{\top} = [A \;\; B]$. It involves, however, the extra step of normalizing the matrix V_2, so that its lower block V_{22} becomes $-I$. This normalization imposes the input/output structure of the model, discussed in Section 1.2, and is the reason for the existence of nongeneric TLS problem. Note that the TLS approximation $[\widehat{A} \;\; \widehat{B}]$ is the same as the low-rank approximation $\widehat{D}^{*\top}$, provided the former exists. Therefore, if one is interested in the best approximation of the data matrix $[A \;\; B]$ and not in the solution X to the system $AX \approx B$, there is no reason to do the normalization of V_2.

We showed that some weighted unstructured low-rank approximation problems have global analytic solution in terms of the SVD. Similar result exists for circulant SLRA. The result is derived independently in the optimization community (Beck & Ben-Tal, 2006) and in the systems and control community (Vanluyten et al., 2005). If the approximation criterion is a unitarily invariant matrix norm, the unstructured low-rank approximation (obtained for example from the truncated SVD) is unique. In the case of a circulant structure, it turns out that this unique minimizer also has circulant structure, so the structure constraint is satisfied without explicitly enforcing it.

An efficient computational way of obtaining the circulant strcutured low-rank approximation is the fast Fourier transform. Consider the scalar case and let

$$P_k := \sum_{j=1}^{n_p} p_j \mathrm{e}^{-(2\pi\mathbf{i}/n_p)kj}$$

be the discrete Fourier transform of p. Denote with $\mathscr{K}$ the subset of $\{1, \ldots, n_p\}$ consisting of the indices of the $\mathtt{m}$ largest elements of $\{|P_1|, \ldots, |P_{n_p}|\}$. Assuming that $\mathscr{K}$ is uniquely defined by the above condition, i.e., assuming that

$$k \in \mathscr{K} \quad \text{and} \quad k' \notin \mathscr{K} \quad \Longrightarrow \quad |P_k| > |P_{k'}|,$$

the solution $\widehat{p}^*$ of the SLRA problem with $\mathscr{S}$ a circulant matrix is unique and is given by

$$\widehat{p}^* = \frac{1}{n_p} \sum_{k \in \mathscr{K}} P_k \mathrm{e}^{(2\pi\mathbf{i}/n_p)\,kj}.$$

4.2. Suboptimal solution methods

The SVD is at the core of many algorithms for approximate modeling, most notably the methods based on balanced model reduction, the subspace identification methods, and the MUSIC and ESPRIT methods in signal processing. The reason for this is that the SVD is a robust and efficient way of computing unstructured low-rank approximation of a matrix. In system identification, signal processing, and computer algebra, however, the low-rank approximation is restricted to the class of matrices with specific (Hankel, Toeplitz, and Sylvester) structure. Ignoring the structure constraint renders the SVD-based methods suboptimal with respect to a desired optimality criterion.

Except for the few special cases described in Section 4.1 there are no global solution methods for general SLRA. The SVD-based methods can be seen as relaxations of the original NP-hard SLRA problem, obtained by removing the structure constraint. Another approach is taken in Fazel (2002), where convex relaxations of the related (see Section 5.3) RMP are proposed. Convex relaxation methods give polynomial time suboptimal solutions.

Presently there is no uniformly best method for computing suboptimal SLRA. In the context of system identification (i.e., block-Hankel SLRA) several subspace and local optimization based methods are compared on practical data sets, see Markovsky, Willems, and De Moor (2006). In general, the heuristic methods are faster but less accurate than the methods based on local optimization, such as the prediction error methods (Ljung, 1999) and the method of Markovsky, Van Huffel, and Pintelon (2005). It is a common practice to use a suboptimal solution obtained by a heuristic method as an initial approximation for an optimization based method. Therefore, the two approaches complement each other.

4.3. Algorithms based on local optimization

Representing the constraint in a kernel form, the SLRA problem becomes the following parameter optimization problem

$$\min_{R,\; RR^{\top}=I_{m-r}} \left(\min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{s.t.} \quad R\mathscr{S}(\widehat{p}) = 0 \right), \tag{20}$$

which is a double minimization problem with a bilinear equality constraint. The outer minimization is over the model parameter R and the inner minimization is over the parameter estimate $\widehat{p}$. Since the mapping $\mathscr{S}$ is affine, there is an affine mapping $\mathscr{G} : \mathbb{R}^{(m-r)\times m} \to \mathbb{R}^{n(m-r)\times n_p}$, such that

$$\operatorname{vec}(R\mathscr{S}(\widehat{p})) = \mathscr{G}(R)\widehat{p} \quad \text{for all } \widehat{p} \in \mathbb{R}^{n_p}. \tag{21}$$

A way to approach the double minimization is by solving the inner minimization analytically, which leads to a nonlinear least squares problem

$$\min_{R,\; RR^{\top}=I_{m-r}} \operatorname{vec}^{\top}\left(R\mathscr{S}(\widehat{p})\right)\left(\mathscr{G}(R)\mathscr{G}^{\top}(R)\right)^{-1}\operatorname{vec}\left(R\mathscr{S}(\widehat{p})\right) \tag{22}$$

for R only (Markovsky, Van Huffel, et al., 2005). The inner minimization problem is a least norm problem and can be given the interpretation of projecting the columns of $\mathscr{S}(p)$ onto the subspace $\mathscr{B} := \ker(R)$, for a given $R \in \mathbb{R}^{m\times(m-r)}$. The projection depends on the parameter R, which is the variable in the outer optimization problem. For this reason, the method is called *variable projections* (Golub & Pereyra, 2003).

In order to evaluate the cost function for the outer minimization problem, we need to solve the inner minimization problem, i.e., the least norm problem $\mathscr{G}(R)z = \operatorname{vec}(R\mathscr{S}(\widehat{p}))$. Direct solution has computational complexity $O(n_p^3)$. The matrix $\mathscr{G}(R)$, however, is structured, which can be used in efficient computational method. The following result from Markovsky, Van Huffel, et al., (2005), shows that for a class of structures $\mathscr{S}$, the structure of the matrix $\mathscr{G}\mathscr{G}^\top$ that appears in the solution of the least norm problem is block-Toeplitz and block-banded.

Theorem 26. *Assume that $\mathscr{S}$ is composed of blocks that are block-Hankel, unstructured, or fixed, i.e.,*

$$\mathscr{S}(p) = [\mathscr{S}_1(p) \cdots \mathscr{S}_q(p)], \tag{23}$$

where $\mathscr{S}_1(p)$ is block-Hankel, unstructured, or does not depend on p. Then the matrix $\mathscr{G}\mathscr{G}^\top$, where $\mathscr{G}$ is defined in (21) *is block-Toeplitz and block-banded structured.*

The implication of Theorem 26 is that for the class of structures (23), efficient $O(n_p)$ cost function evaluation can be done by Cholesky factorization of a block-Toeplitz banded matrix. The SLICOT library includes high quality FORTRAN implementation of algorithms for this problem. It is used in a software package for solving SLRA problems, based on the Levenberg–Marquardt algorithm, implemented in MINPACK (Markovsky, Van Huffel, et al., 2005; Markovsky & Van Huffel, 2005). This algorithm is globally convergent with a superlinear convergence rate.

Some SLRA problems can be solved by an algorithm based on the alternating least squares method. Consider the approximate deconvolution problems

$$\min_{\widehat{u},\widehat{y},h} \quad \|\operatorname{col}(u_\mathrm{d}, y_\mathrm{d}) - \operatorname{col}(\widehat{u}, \widehat{y})\|_2$$
$$\text{s.t.} \quad h^\top \mathscr{T}_{\ell+1}(\widehat{u}) = \widehat{y}^\top. \tag{24}$$

It is equivalent to the problem

$$\min_{\widehat{u},h} \left\| \begin{bmatrix} u_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} - \begin{bmatrix} \widehat{u} \\ \mathscr{T}_{\ell+1}^\top(\widehat{u})h \end{bmatrix} \right\|_2. \tag{25}$$

The alternating projections algorithm, see Algorithm 1, is based on the fact that problem (25) is a standard least squares problem for given $\widehat{u}$ and for given h. Minimizing alternatively over h with $\widehat{u}$ fixed to its value from the previous iteration step, and over $\widehat{u}$ with h fixed its value from the previous iteration step, we obtain a sequence of approximations $(\widehat{u}^{(k)}, h^{(k)})$, $k = 1, 2, \ldots$ that corresponds to a nonincreasing sequence of cost function values. The alternating projections algorithm is also globally convergent, however, its local convergence rate is only linear.

In this section we described the variable projections and alternating projections methods for solving the SLRA problem. Using global optimization methods, e.g., the branch and bound type algorithms, instead of local optimization methods is also an option. Efficient cost function evaluation, obtained by exploiting the matrix structure, is of prime importance in the application of global optimization methods as well. The number of cost function evaluations required for finding a global solution, however, is likely to be much higher than the one required for finding a locally optimal solution.

Algorithm 1. Alternating projections algorithm for solving the approximate deconvolution problem (24).

```
Input: Data $u_d$, $y_d$ and convergence tolerance $\varepsilon$.
Set $k := 0$ and compute an initial approximation $h^{(0)}, \widehat{u}^{(0)}, \widehat{y}^{(0)}$, e.g., by solving the TLS problem $\mathscr{T}^\top(u_\mathrm{d})h = y_\mathrm{d}$.
repeat
    $k := k + 1$.
    Solve the least squares problem in $\widehat{u}$
        $\widehat{u}^{(k)} := \arg\min_{\widehat{u}} \left\| \begin{bmatrix} u_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} - \begin{bmatrix} I \\ \widetilde{\mathscr{T}}\ (h^{(k-1)}) \end{bmatrix} \widehat{u} \right\|_2$,
        where $\widetilde{\mathscr{T}}\ (h)$ is a matrix depending on $h$, such that $\mathscr{T}_{\ell+1}^\top(\widehat{u})h = \widetilde{\mathscr{T}}\ (h)\widehat{u}$.
    Solve the least squares problem in $h$
        $h^{(k)} := \arg\min_h \|y_\mathrm{d} - \mathscr{T}_{\ell+1}^\top(\widehat{u}^{(k)})h\|_2$.
    $\widehat{y}^{(k)} := \mathscr{T}_{\ell+1}^\top(\widehat{u}^{(k)})h^{(k)}$.
until $\|\operatorname{col}(\widehat{u}^{(k-1)}, \widehat{y}^{(k-1)}) - \operatorname{col}(\widehat{u}^{(k)}, \widehat{y}^{(k)})\| < \varepsilon$
Output: A locally optimal solution $h^* := h^{(k)}$, $\widehat{u}^* := \widehat{u}^{(k)}$, and $\widehat{y}^* := \widehat{y}^{(k)}$ of (24).
```

4.4. Simulation example

The database for identification of systems (DAISY) (De Moor, 2005) contains real-life and simulated benchmark data sets. In this section, we use the DAISY data set called "Wing flutter data", which consist of $T = 1024$ samples of the input and the output of the system to be identified.[2] We divide the data $w_\mathrm{d} = (u_\mathrm{d}, y_\mathrm{d})$ into identification $w_\mathrm{idt} = (u_\mathrm{idt}, y_\mathrm{idt})$ and validation $w_\mathrm{val} = (u_\mathrm{val}, y_\mathrm{val})$ parts, see Fig. 3. An optimal model $\widehat{\mathscr{B}}^*$ in the sense of Problem 5 is identified from w_idt using Proposition 6 and the software package of (Markovsky & Van Huffel, 2005) and is validated on w_val, using the fitting error

$$e(w_\mathrm{val}, \widehat{\mathscr{B}}^*) := \min_{\widehat{w}^* \in \widehat{\mathscr{B}}^*} \|w_\mathrm{val} - \widehat{w}^*\|. \tag{26}$$

The optimal model is searched in the LTI model class $\mathscr{L}_{\mathtt{m},\ell}$, where $\mathtt{m} = 1$ is fixed by the given data but ℓ is an unknown parameter. We choose ℓ from the fitting error $e(w_\mathrm{idt}, \widehat{\mathscr{B}}^*)$ vs complexity ℓ trade-off curve, see Fig. 4. The "best" value for the parameter ℓ is selected visually and corresponds to the corner of the "L" shaped trade-off curve. In the particular example, we choose $\ell = 4$, so the considered model class is $\mathscr{L}_{1,4}$. The optimal model $\widehat{\mathscr{B}}^*$ in $\mathscr{L}_{1,4}$ is $\widehat{\mathscr{B}}^* = \ker(\widehat{R}^*(\sigma))$, where

$$\begin{aligned}\widehat{R}^*(z) = {} & [9.55 \quad -0.09]z^0 + [-32.18 \quad 1.50]z^1 \\ & + [43.77 \quad -3.56]z^2 + [-28.62 \quad 3.05]z^3 \\ & + [7.57 \quad -1]z^4.\end{aligned}$$

[2] The description of this data set in DAISY says "Due to industrial secrecy agreements we are not allowed to reveal more details."

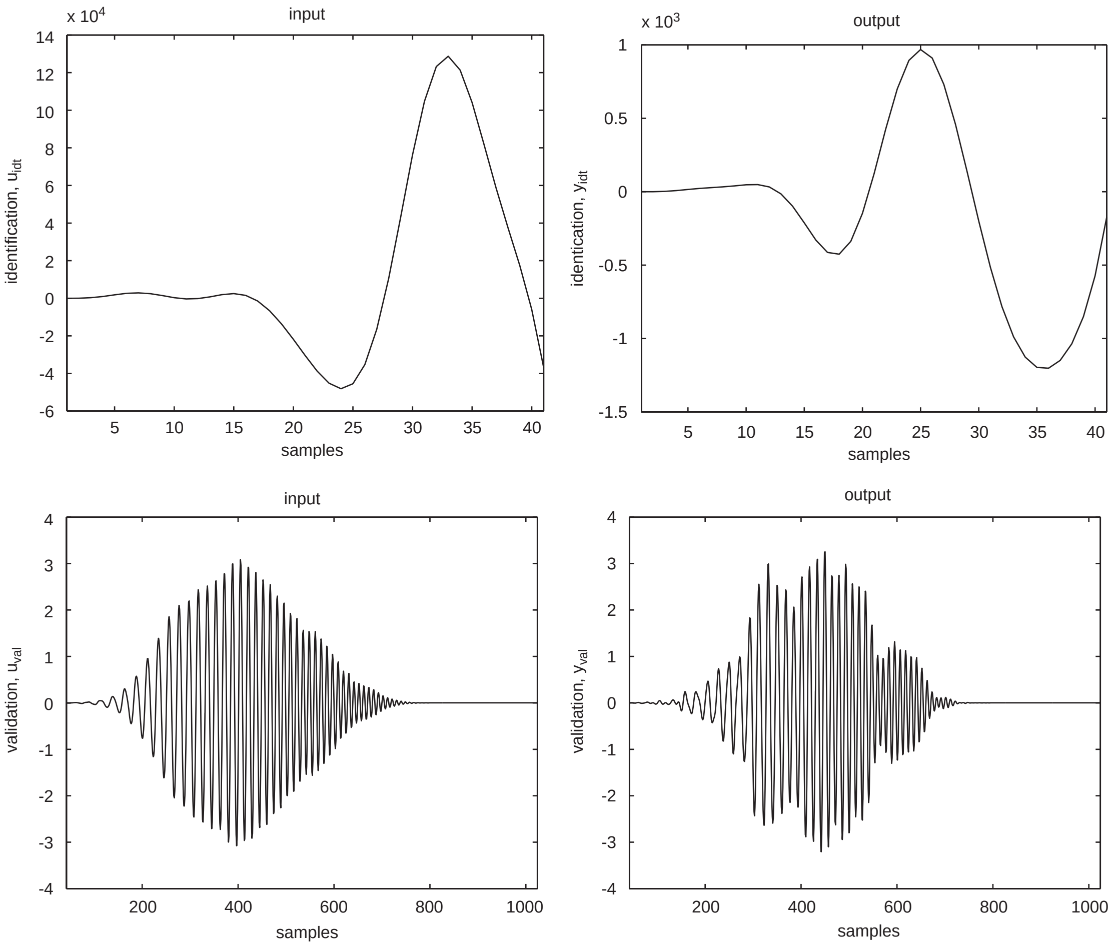

Fig. 3. Identification $(u_{\text{idt}}, y_{\text{idt}})$ and validation $(u_{\text{val}}, y_{\text{val}})$ parts of the wing flutter data.

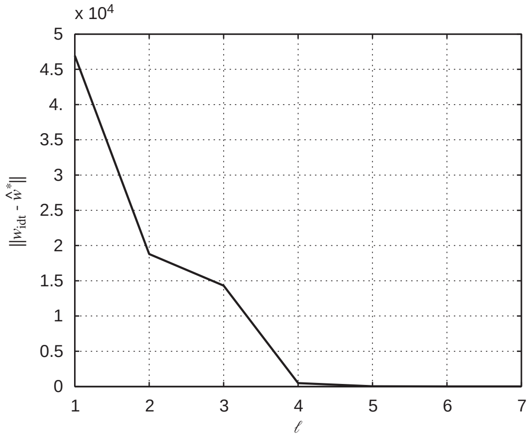

Fig. 4. Misfit vs complexity trade-off curve.

The validation of $\widehat{\mathscr{B}}^*$ on w_{val} is shown in Fig. 5. A relatively simple (fourth order) model explains accurately the validation part of the data in the sense of the error criterion (26). Note that the identification and the validation criteria are defined in the errors-in-variables setting, i.e., both the input and the output are assumed perturbed and are modified. The choice of the errors-in-variables setup in the wing flutter example is ad hoc, however, the simulation results suggest that it is an adequate choice for the wing flutter example, because from an extremely small portion of the data (only 4%) the identified model can explain sufficiently well the remaining part of the data.

5. Generalizations and related problems

In the previous sections we reviewed the structured low-rank approximation problem and some of its applications. In this section, we describe other approximation problems that generalize the basic unstructured low-rank approximation problem by considering alternative cost functions and extra constraints on the approximant. We review also the related problems of rank minimization and structured pseudospectra.

5.1. Weighted low-rank approximation

The motivation for the WLRA problem

$$\min_{\widehat{D}} \ \operatorname{vec}^\top(D - \widehat{D})W^{-1}\operatorname{vec}(D - \widehat{D}) \quad \text{s.t.} \quad \operatorname{rank}(\widehat{D}) \leqslant \mathtt{m} \qquad \text{(WLRA)}$$

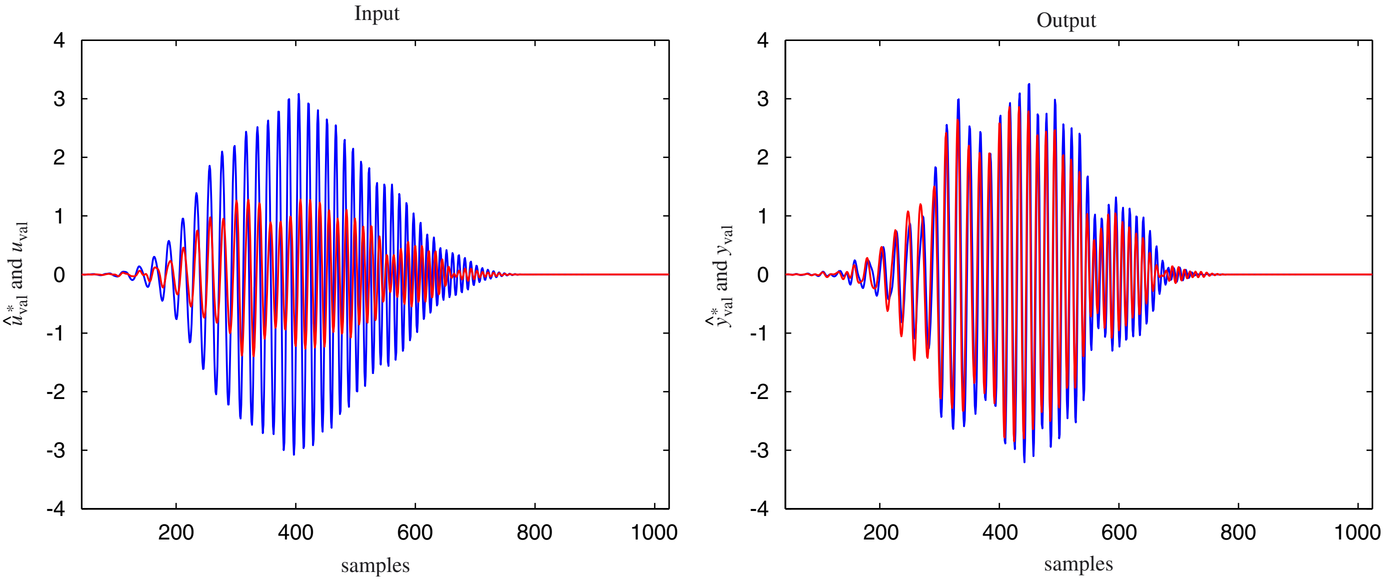

Fig. 5. Validation of the model $\widehat{\mathscr{B}}^*$ on the data w_{val}.

is to compute the maximum likelihood estimator in the errors-in-variables setting, see Proposition 3. Similar to the SLRA problem, the WLRA problem is, in general, NP-hard and methods based on heuristics, local, and global optimization methods are used for its solution.

The heuristic methods in this case are based on replacement of the weight matrix W by a matrix $\widehat{W}$ of the form $\widehat{W} = W_{\text{r}} \otimes W_{\text{l}}$, where $W_{\text{r}} \in \mathbb{R}^{N\times N}$ and $W_{\text{l}} \in \mathbb{R}^{\mathtt{d}\times\mathtt{d}}$ are positive definite matrices and $\otimes$ is the Kronecker product. The reason for this is that the WLRA problem

$$\min_{\widehat{D}} \operatorname{vec}^\top(D-\widehat{D})(W_{\text{r}}\otimes W_{\text{l}})^{-1}\operatorname{vec}(D-\widehat{D}) \qquad \text{s.t.}\quad \operatorname{rank}\ (\widehat{D})\leqslant \mathtt{m} \tag{WLRA$'$}$$

has an analytic solution in terms of the SVD.

Theorem 27. *Define the modified data matrix*

$$D_{\text{m}} := \left(\sqrt{W_{\text{l}}}\right)^{-1} D \left(\sqrt{W_{\text{r}}}\right)^{-1},$$

and let $\widehat{D}^*_{\text{m}}$ be the optimal (unweighted) low rank approximation of D_{m}. Then

$$\widehat{D}^* := \sqrt{W_{\text{l}}}\widehat{D}^*_{\text{m}}\sqrt{W_{\text{r}}},$$

is a solution of problem (WLRA$'$). A solution always exists. It is unique if and only if $\widehat{D}^*_{\text{m}}$ is unique.

Using the kernel representation, problem (WLRA) becomes

$$\min_{\widehat{D},R} \operatorname{vec}^\top(D-\widehat{D})W^{-1}\operatorname{vec}(D-\widehat{D}) \qquad \text{s.t.}\quad R\widehat{D}=0 \quad \text{and} \quad RR^\top = I_{\mathtt{d}-\mathtt{m}}. \tag{27}$$

A class of methods, following the variable projections approach, is described in Manton, Mahony, and Hua (2003) and Markovsky and Van Huffel (2007). The equivalent optimization problem obtained by eliminating the variable $\widehat{D}$ is

$$\min_{R,\ RR^\top = I_{\mathtt{d}-\mathtt{m}}} \operatorname{vec}^\top(D)(I_N\otimes R)^\top\left((I_N\otimes R)\ W\ (I_N\otimes R)^\top\ \right)^{-1} \times (I_N\otimes R)\operatorname{vec}(D).$$

This problem is treated in Manton et al. (2003) as an optimization on a Grassman manifold (set of matrices with a certain specified rank) and a new class of local optimization methods is derived. An alternative approach that is based on the alternating projections method is popular in chemometrics, where the problem is known as the maximum likelihood principal component analysis (Wentzell, Andrews, Hamilton, Faber, & Kowalski, 1997).

5.2. Nonnegative low-rank approximation

We use the notation $D\geqslant 0$ for a matrix $D\in\mathbb{R}^{\mathtt{d}\times N}$ whose elements are nonnegative. A low-rank approximation problem with elementwise nonnegativity constraint

$$\min_{\widehat{D}} \quad \|D-\widehat{D}\| \qquad \text{s.t.}\quad \operatorname{rank}(\widehat{D})\leqslant r \quad \text{and} \quad \widehat{D}\geqslant 0 \tag{28}$$

arises in Markov chains (Vanluyten, Willems, & De Moor, 2006) and image mining (Lee & Seung, 1999).

Using the image representation, we obtain the following problem:

$$\min_{\substack{\widehat{D},\ P\in\mathbb{R}^{\mathtt{d}\times\mathtt{m}}\\ L\in\mathbb{R}^{\mathtt{m}\times N}}} \quad \|D-\widehat{D}\| \qquad \text{s.t.}\quad \widehat{D}=PL \quad \text{and} \quad P,L\geqslant 0, \tag{29}$$

which is a relaxation of problem (28). The minimal inner dimension $\mathtt{m}$ in the factorization $\widehat{D} = PL$, where P and L are elementwise nonnegative is called the *positive rank* of $\widehat{D}$ (Berman & Shaked-Monderer, 2003). In general, the positive rank is less than or equal to the rank.

Note that due to the nonnegativity constraint on $\widehat{D}$, the problem cannot be solved using the variable projections method. (There is no closed form solution for the equivalent problem with $\widehat{D}$ eliminated.) The alternating projections algorithm, however, can be used almost without modification for the solution of the relaxed problem (29). Let the norm $\|\cdot\|$ in (28) be the Frobenius norm.[3] Then on each iteration step of the algorithm two least squares problems with nonnegativity constraint (i.e., standard optimization problems) are solved. The resulting alternating least squares algorithm is Algorithm 2.

Algorithm 2. Alternating projections algorithm for nonnegative low-rank approximation.

```
Input: Data matrix $D$, desired rank $\mathtt{m}$, and convergence tolerance $\varepsilon$.
Set $k := 0$ and compute an initial approximation $\widehat{D}^{(0)} := P^{(0)}L^{(0)}$ from the SVD by setting all negative elements to zero.
repeat
    $k := k + 1$.
    Solve: $L^{(k)} := \arg\min_L \|D - P^{(k-1)}L\|$ s.t. $L \geqslant 0$.
    Solve: $P^{(k)} := \arg\min_P \|D - PL^{(k)}\|$ s.t. $P \geqslant 0$.
until $\|P^{(k-1)}L^{(k-1)} - P^{(k)}L^{(k)}\| < \varepsilon$
Output: A locally optimal solution $\widehat{D}^* := P^{(k)}L^{(k)}$ to problem (29).
```

5.3. Rank minimization

Approximate modeling is a trade-off between fitting accuracy $\|p - \widehat{p}\|$ and model complexity $\operatorname{rank}(\mathscr{S}(\widehat{p}))$. Two possible scalarizations of the bi-objective optimization are:

- *low-rank approximation*: maximizing the fitting accuracy under a constraint r on the complexity, and
- *rank minimization*: minimize the complexity under a constraint γ on the fitting accuracy, i.e.,

$$\min_{\widehat{p}} \quad \operatorname{rank}(\mathscr{S}(\widehat{p})) \qquad \text{s.t.} \quad \|p - \widehat{p}\| \leqslant \gamma. \tag{30}$$

The optimal cost function values and corresponding constraint levels of both problems describe the same trade-off curve of Pareto optimal solutions. Therefore, an algorithm for solving the RMP can solve the low-rank approximation problem by using bisection.

In Fazel (2002), the following more general formulation of the RMP is considered:

$$\min_{\widehat{D}} \quad \operatorname{rank}(\widehat{D}) \qquad \text{s.t.} \quad \widehat{D} \in \mathscr{C}, \tag{RMP}$$

where $\mathscr{C}$ is a convex set. Note that the convex constraint $\widehat{D} \in \mathscr{C}$ is much more general than the norm constraint $\|p - \widehat{p}\| \leqslant \gamma$. For example, the former can impose in addition to a fitting criterion $\|D - \widehat{D}\| \leqslant \gamma$ element-wise positivity of $\widehat{D}$, cf., Section 5.2.

Heuristic methods for solving the RMP are presented in Fazel (2002). Consider for simplicity the case when $\widehat{D}$ is positive definite. By replacing the rank function with the trace function, we obtain a convex relaxation of (RMP). The rationale behind this heuristic is that minimization of the trace of a positive definite matrix tends to minimize its rank (more precisely the smallest singular values tend to be close to zero). Note that for $\widehat{D} \succ 0$ and $\widehat{D}$ diagonal, the trace heuristic corresponds to the popular ℓ_1 heuristic for obtaining sparse solutions.

A number of applications of the RMP are presented in Fazel (2002). Among them are system realization with time domain constraints, reduced order controller design, and frequency domain system approximation.

5.4. Structured pseudospectra

Let $\Lambda(A)$ be the set of eigenvalues of $A \in \mathbb{C}^{n \times n}$, and $\mathbb{M}$ be the set of structured matrices $\mathbb{M} := \{\mathscr{S}(p) \mid p \in \mathbb{R}^{n_p}\}$, with a given structure specification $\mathscr{S}$. The structured pseudospectra (Graillat, 2006; Trefethen & Embree, 1999) of A is defined as follows:

$$\Lambda_\varepsilon(A) := \{z \in \mathbb{C} \mid z \in \Lambda(\widehat{A}),\ \widehat{A} \in \mathbb{M},\ \|A - \widehat{A}\| \leqslant \varepsilon\}.$$

Using, $\Lambda_\varepsilon(A)$ one can determine the structured distance of A to singularity

$$s(A) := \min_{\Delta A} \|\Delta A\| \qquad \text{s.t.} \quad A + \Delta A \text{ is singular and } \Delta A \in \mathbb{M}.$$

This is a special SLRA problem for square data matrix and rank reduction by 1. Closely related to the structured pseudospectra problem is the structured condition number problem for a linear system of equations, see Rump (2003).

6. Conclusions

We reviewed the SLRA problem from the data fitting point of view. The abstract rank constraint is related to the existence of a linear model that fits the data. If, in addition to being low-rank, the data matrix is Hankel structured, then the fitting model, in addition to being linear, is time-invariant dynamic. In the special case of unstructured low-rank approximation the model is static. A commonly used method to achieve a low-rank approximation is to solve approximately an overdetermined linear system of equations, e.g., in the total least squares sense. This approach,

[3] In the context of Markov chains more adequate is the choice of the Kullback–Leibler divergence as a distance measure between D and $\widehat{D}$.

however, imposes an additional input/output structure on the model that might not be relevant in the application at hand.

There are numerous applications of SLRA in system theory, signal processing, and computer algebra. The data matrix is block structured where each of the blocks is either block-Hankel, unstructured, or fixed. The model being multivariable implies that the data matrix is block-Hankel structured with unstructured block elements. The model being multidimensional implies that the data matrix is block-Hankel structured with Hankel structured block elements. We reviewed algorithms for solving low-rank approximation problems, based on the variable projections and alternating projections methods.

Finally, we showed generalizations and related problems. The generalizations consider different approximation criteria and constraints on the data matrix. Closely related to structured low-rank minimization are the rank minimization and the structured pseudo-spectra problems.

Appendix A. Kernel, image, and input/output representations

A.1. Linear static models

A static model $\mathscr{B}$ with $\mathtt{d}$ variables is a subset of $\mathbb{R}^{\mathtt{d}}$. A linear static model with $\mathtt{d}$ variables is a subspace of $\mathbb{R}^{\mathtt{d}}$. Three basic representations of a linear static model $\mathscr{B} \subseteq \mathbb{R}^{\mathtt{d}}$ are the kernel, image, and input/output ones:

- kernel representation $\mathscr{B} = \ker(R)$, with $R \in \mathbb{R}^{\mathtt{p}\times\mathtt{d}}$,
- image representation $\mathscr{B} = \operatorname{image}(P)$, with $P \in \mathbb{R}^{\mathtt{d}\times\mathtt{m}}$, and
- input/output representation

$$\mathscr{B}_{\mathrm{i/o}}(X, \Pi) := \{d := \Pi \operatorname{col}(d_{\mathrm{i}}, d_{\mathrm{o}}) \in \mathbb{R}^{\mathtt{d}} \mid d_{\mathrm{i}} \in \mathbb{R}^{\mathtt{m}}, d_{\mathrm{o}} = X^\top d_{\mathrm{i}}\}$$

with parameters $X \in \mathbb{R}^{\mathtt{m}\times\mathtt{p}}$ and a permutation matrix Π.

If the parameter Π in an input/output representation is not specified, then by default it is $\Pi = I$, i.e., the first $\mathtt{m}$ variables are assumed inputs and the other $\mathtt{p} := \mathtt{d}-\mathtt{m}$ outputs. In terms of the data matrix D, the input/output representation is the system of equations $AX \approx B$, where $[A \;\; B]\Pi^\top := D^\top$. Thus solving $AX \approx B$ approximately by least squares, TLS, or any other method is equivalent to solving a low-rank approximation using an input/output representation.

If the parameters R, P, and X describe the same system $\mathscr{B}$, then they are related. Let $\Pi = I$ and define the partitioning

$$R := [R_{\mathrm{i}} \;\; R_{\mathrm{o}}] \quad \text{where } R_{\mathrm{o}} \in \mathbb{R}^{\mathtt{p}\times\mathtt{p}}$$

and

$$P := \begin{bmatrix} P_{\mathrm{i}} \\ P_{\mathrm{o}} \end{bmatrix} \quad \text{where } P_{\mathrm{i}} \in \mathbb{R}^{\mathtt{m}\times\mathtt{m}}.$$

Fig. A.1 shows the links among the parameters R, P, and X.

$$\begin{tikzcd}
\mathscr{B}=\ker(R) \arrow[rr, leftrightarrow, "RP=0"] \arrow[dr, shift left, "X^\top=-R_{\mathrm{o}}^{-1}R_{\mathrm{i}}"] & & \mathscr{B}=\operatorname{image}(P) \arrow[dl, shift left, "X^\top=P_{\mathrm{o}}P_{\mathrm{i}}^{-1}"] \\
& \mathscr{B}=\mathscr{B}_{\mathrm{i/o}}(X) \arrow[ul, shift left, "R=[X^\top \;\; -I]"] \arrow[ur, shift left, "P^\top=[I \;\; X]"] &
\end{tikzcd}$$

Fig. A.1. Links among the parameters R, P, and X.

In an image representation $\operatorname{image}(P) = \mathscr{B}$, the columns of P are *generators* of the subspace $\mathscr{B}$. In a kernel representation $R\mathscr{B} = 0$, the rows of R are *annihilators* of $\mathscr{B}$. The parameters R and P are not unique due to:

1. non-minimality of the set of annihilators/generators of $\mathscr{B}$, and
2. change of basis: $\ker(R) = \ker(UR)$, for all U, $\det(U) \neq 0$; and $\operatorname{image}(P) = \operatorname{image}(PV)$, for all V, $\det(V) \neq 0$.

The smallest possible col dim(P), such that $\operatorname{image}(P) = \mathscr{B}$ is $\mathtt{m} := \dim(\mathscr{B})$—the number of inputs of $\mathscr{B}$. With col dim(P)=$\mathtt{m}$, the columns of P form a basis for $\mathscr{B}$. The smallest possible row dim(R), such that $\ker(R) = \mathscr{B}$ is $\mathtt{p} := \mathtt{d} - \mathtt{m}$—the number of outputs of $\mathscr{B}$. With row dim(R) $= \mathtt{p}$, the rows of R form a basis for the orthogonal complement $\mathscr{B}^\perp$ of $\mathscr{B}$. Therefore, without loss of generality we can assume that $P \in \mathbb{R}^{\mathtt{d}\times\mathtt{m}}$ and $R \in \mathbb{R}^{\mathtt{p}\times\mathtt{d}}$.

In general, many input/output partitions of the variables d are possible. Choosing an input/output partition amounts to choosing a full rank $\mathtt{p} \times \mathtt{p}$ submatrix of R or a full rank $\mathtt{m} \times \mathtt{m}$ submatrix of P. Often there is no a priori reason to prefer one partition over another. Thus $AX \approx B$ is often not a natural starting point for data modeling.

A.2. Linear time-invariant dynamic models

A discrete-time dynamic model $\mathscr{B}$ with $\mathtt{w}$ variables is a subset of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$, the set of all functions from the time axis $\mathbb{N} := \{1, 2, \ldots\}$ to the variable space $\mathbb{R}^{\mathtt{w}}$. By definition, a model $\mathscr{B}$ is LTI if $\mathscr{B}$ is a shift-invariant subspace of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{N}}$. If in addition, $\mathscr{B}$ is a closed subset (in the topology of point-wise convergence), then it turns out that $\mathscr{B}$ is finite-dimensional. This means that at any time t the future behavior of the system is deterministically independent of the past behavior, given a finite dimensional vector, called a state of the system. The smallest state dimension is an important invariant of the system and is called the order. Another invariant of the system that we will often use is the lag (also called the observability index). The lag of a finite dimensional LTI system is the smallest degree of a difference equation representation of the system.

Let $\mathscr{B}$ be LTI with $\mathtt{m}$ inputs, $\mathtt{p}$ outputs, of order $\mathtt{n}$, and lag ℓ, then the restriction $\mathscr{B}_{\overline{[1,T]}}$ of $\mathscr{B}$ to the interval $[1, T]$ has dimension

$$\dim(\mathscr{B}_{\overline{[1,T]}}) \leqslant \mathtt{m}T + \mathtt{n} \leqslant \mathtt{m}T + \mathtt{p}\ell \quad \text{for } T \geqslant \ell.$$

The number $\dim(\mathscr{B}_{\overline{[1,T]}})$ is a measure of the model complexity: the larger $\dim(\mathscr{B})$ is, the more complicated the model is. Therefore, the complexity of $\mathscr{B}$ can be specified by the pair

$(\mathtt{m}, \mathtt{n})$ or alternatively by the pair $(\mathtt{m}, \ell)$. We use the notation $\mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell}$ for the LTI model class with bounded complexity: $\mathtt{m}$ or less inputs and lag at most ℓ.

Three common representations for LTI model are:

- kernel representation

$$R_0 w(t) + R_1 w(t+1) + \cdots + R_\ell w(t+\ell) = 0,$$

  with parameter the polynomial matrix $R(z) := \sum_{i=0}^{\ell} R_i z^i$,
- impulse response representation

$$w = \Pi \operatorname{col}(u, y), \qquad y(t) = \sum_{\tau=0}^{\infty} H(\tau) u(t-\tau),$$

  with parameters the signal $H : \mathbb{N} \to \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$ and permutation matrix Π, and
- input/state/output representation

$$w = \Pi \operatorname{col}(u, y), \qquad \begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

  with parameters (A, B, C, D) and permutation matrix Π.

Transitions among the parameters R, H, and (A, B, C, D) are classical problems, e.g., going from R or H to (A, B, C, D) are realization problems.

Appendix B. Proofs

B.1. Proof of Proposition 3

The probability density function of the observation vector $\operatorname{vec}(D)$ is

$$p_{\widehat{\mathscr{B}},\widehat{D}}(\operatorname{vec}(D)) = \begin{cases} \text{const} \cdot \exp -\dfrac{1}{2v} \|\operatorname{vec}(D) - \operatorname{vec}(\widehat{D})\|_W^2, \\ \qquad \text{if } \operatorname{image}(\widehat{D}) \subset \widehat{\mathscr{B}} \text{ and } \dim(\widehat{\mathscr{B}}) \leqslant \mathtt{m} \\ 0, \qquad \text{otherwise,} \end{cases}$$

where "const" is a term that does not depend on $\widehat{D}$ and $\widehat{\mathscr{B}}$. The log-likelihood function is

$$\ell(\widehat{\mathscr{B}}, \widehat{D}) = \begin{cases} \text{const} \quad -\dfrac{1}{2v} \|\operatorname{vec}(D) - \operatorname{vec}(\widehat{D})\|_W^2 \\ \qquad \text{if } \operatorname{image}(\widehat{D}) \subset \widehat{\mathscr{B}} \text{ and } \dim(\widehat{\mathscr{B}}) \leqslant \mathtt{m} \\ -\infty \quad \text{otherwise,} \end{cases}$$

and the maximum likelihood estimation problem is

$$\min_{\widehat{\mathscr{B}},\widehat{D}} \quad \frac{1}{2v} \|\operatorname{vec}(D) - \operatorname{vec}(\widehat{D})\|_W^2$$
$$\text{s.t.} \quad \operatorname{image}(\widehat{D}) \subset \widehat{\mathscr{B}} \quad \text{and} \dim(\widehat{\mathscr{B}}) \leqslant \mathtt{m},$$

which is an equivalent problem to Problem 2 with $\|\cdot\| = \|\cdot\|_W$.

Remark 28. An image representation $\operatorname{image}(\widehat{P}^*)$ of the optimal approximate model $\widehat{\mathscr{B}}^*$ can be obtained from the solution $\widehat{D}^*$ to the low-rank approximation problem by doing a rank revealing factorization $\widehat{D}^* = \widehat{P}^* \widehat{L}^*$, and a kernel representation $\ker(\widehat{R}^*)$ can be obtained by computing a basis for the left nullspace of $\widehat{D}^*$. The parameters $\widehat{P}$ and $\widehat{R}$, however, can be used as optimization variables in the low-rank approximation problem, in which case they are obtained as a by-product from the algorithm solving the low-rank approximation problem.

B.2. Proof of Proposition 6

It is analogous to the proof of Proposition 3 and is skipped.

Remark 29. Computing the optimal approximate model $\widehat{\mathscr{B}}^*$ from the solution $\widehat{p}^*$ to the SLRA problem is an exact identification problem, see Markovsky, Willems, Van Huffel et al. (2006, Chapter 8), Van Overschee and De Moor (1996, Chapter 2), and Markovsky, Willems, Rapisarda, and De Moor (2005). As in the static approximation problem, however, the parameter of a model representation is an optimization variable of the optimization problem, used for solving the SLRA problem, so that a representation of the model is obtain directly from the optimization solver.

B.3. Proof of Proposition 6

The proposition follows from the equivalence

$$\widehat{H} \text{ is the impulse response of } \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$$
$$\iff \quad \operatorname{rank}(\mathscr{H}_{\ell+1}(\sigma\widehat{H})) \leqslant \mathtt{m}\ell,$$

which is a well-known fact from realization theory, see, e.g., Markovsky, Willems, Van Huffel et al. (2006, Section 8.7). Obtaining the model $\widehat{\mathscr{B}}$ from the solution to the SLRA problem is the LTI system realization problem.

B.4. Proof of Proposition 10

The problem is equivalent to the approximate realization problem with H_{d} being the impulse response of $\mathscr{B}_{\mathrm{d}}$, see Markovsky, Willems, Van Huffel et al. (2006, Section 11.4).

B.5. Proof of Proposition 12

The proposition is based on the equivalence

$$(u_{\mathrm{d}}, \widehat{y}) \in \widehat{\mathscr{B}}_{[1,T]} \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$$
$$\iff \quad \operatorname{rank}\left(\begin{bmatrix} \mathscr{H}_{\ell+1}(u_{\mathrm{d}}) \\ \mathscr{H}_{\ell+1}(\widehat{y}) \end{bmatrix}\right) \leqslant \mathtt{m}(\ell+1) + \mathtt{p}\ell,$$

which is a corollary of the following lemma.

Lemma 30. *The signal $w \in \mathscr{B}_{[1,T]}$ and $\mathscr{B} \in \mathscr{L}^{\mathtt{w}}_{\mathtt{m},\ell}$ if and only if* $\operatorname{rank}(\mathscr{H}_{\ell+1}(w)) \leqslant \mathtt{m}(\ell+1) + (\mathtt{w} - \mathtt{m})\ell$.

Proof. (⇒) Assume that $w \in \mathscr{B}_{\overline{[1,T]}}$, $\mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell}$ and let $\mathscr{B}(R)$, with $R(z)=\sum_{i=0}^{\ell} z^i R_i \in \mathbb{R}^{g\times\mathtt{w}}[z]$ full row rank, be a kernel representation of $\mathscr{B}$. The assumption $\mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell}$ implies that $g \geqslant \mathtt{p} := \mathtt{w}-\mathtt{m}$. From $w \in \mathscr{B}_{\overline{[1,T]}}$, we have that

$$[R_0\ \ R_1\ \ \cdots\ \ R_\ell]\mathscr{H}_{\ell+1}(w)=0,$$

which implies that $\operatorname{rank}(\mathscr{H}_{\ell+1}(w)) \leqslant \mathtt{m}(\ell+1)+\mathtt{p}\ell$.

(⇐) Assume that $\operatorname{rank}(\mathscr{H}_{\ell+1}(w)) \leqslant \mathtt{m}(\ell+1)+\mathtt{p}\ell$. Then there is a full row rank matrix $R := [R_0\ \ R_1\ \ \cdots\ \ R_\ell] \in \mathbb{R}^{\mathtt{p}\times\mathtt{w}(\ell+1)}$, such that $R\mathscr{H}_{\ell+1}(w)=0$. Define the polynomial matrix $R(z)=\sum_{i=0}^{\ell} z^i R_i$. Then the system $\mathscr{B}$ induced by $R(z)$ via the kernel representation (13) is such that $w \in \mathscr{B}$ and $\mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell}$. □

B.6. Proof of Proposition 16

The problem is equivalent to the approximate realization problem, see Markovsky, Willems, Van Huffel et al. (2006, Section 11.4).

B.7. Proof of Proposition 18

The proposition follows from the equivalence

$$\widehat{w} \in \mathscr{B}_{\overline{[1,T]}} \text{ and } \widehat{\mathscr{B}} \in \mathrm{FIR}_{\mathtt{m},\ell} \iff \operatorname{rank}\left(\begin{bmatrix} [\widehat{y}(1)\ \ \cdots\ \ \widehat{y}(T-\ell)] \\ \mathscr{H}_{\ell+1}(\widehat{u}) \end{bmatrix}\right) \leqslant \mathtt{m}(\ell+1).$$

In order to show it, let

$$H=(H(0),H(1),\ldots,H(\ell),0,0,\ldots)$$

be the impulse response of $\widehat{\mathscr{B}} \in \mathrm{FIR}_{\mathtt{m},\ell}$. The signal $\widehat{w}=(\widehat{u},\widehat{y})$ is a trajectory of $\mathscr{B}$ if and only if

$$[H(\ell)\ \ \cdots\ \ H(1)\ \ H(0)]\mathscr{H}_{\ell+1}(\widehat{u})=[\widehat{y}(1)\ \ \cdots\ \ \widehat{y}(T-\ell)]$$

or equivalently

$$[-I_g\ \ H(\ell)\ \ \cdots\ \ H(1)\ \ H(0)]\begin{bmatrix} [\widehat{y}(1)\ \ \cdots\ \ \widehat{y}(T-\ell)] \\ \mathscr{H}_{\ell+1}(\widehat{u}) \end{bmatrix}=0.$$

The assumption $\widehat{\mathscr{B}} \in \mathrm{FIR}_{\mathtt{m},\ell}$ implies that $g \geqslant \mathtt{p}$. Therefore,

$$\operatorname{rank}\left(\begin{bmatrix} [\widehat{y}(1)\ \ \cdots\ \ \widehat{y}(T-\ell)] \\ \mathscr{H}_{\ell+1}(\widehat{u}) \end{bmatrix}\right) \leqslant \mathtt{m}(\ell+1).$$

B.8. Proof of Proposition 21

The proposition follows from the equivalence

$$\widehat{y} \in \widehat{\mathscr{B}}_{\overline{[1,T]}},\ \ \widehat{\mathscr{B}} \in \mathscr{L}^1_{0,\ell} \text{ and } \widehat{\mathscr{B}}(\widehat{R})=\widehat{\mathscr{B}} \text{ is palindromic} \iff \operatorname{rank}\left([\mathscr{H}_{\ell+1}(y)\ \ \mathscr{T}_{\ell+1}(y)]\right) \leqslant \ell.$$

In order to show it, let $\mathscr{B}(R)$, with $R(z)=\sum_{i=0}^{\ell} z^i R_i$ full row rank, be a kernel representation of $\mathscr{B} \in \mathscr{L}^1_{0,\ell}$. Then $\widehat{y} \in \widehat{\mathscr{B}}_{\overline{[1,T]}}$ is equivalent to $[R_0\ \ R_1\ \ \cdots\ \ R_\ell]\mathscr{H}_{\ell+1}(\widehat{y})=0$. If, in addition, R is palindromic, then

$$[R_\ell\ \ \cdots\ \ R_1\ \ R_0]\mathscr{H}_{\ell+1}(\widehat{y})=0 \iff [R_0\ \ R_1\ \ \cdots\ \ R_\ell]\mathscr{T}_{\ell+1}(\widehat{y})=0.$$

We have that

$$[R_0\ \ R_1\ \ \cdots\ \ R_\ell][\mathscr{H}_{\ell+1}(y)\ \ \mathscr{T}_{\ell+1}(y)]=0. \tag{B.1}$$

which is equivalent to $\operatorname{rank}([\mathscr{H}_{\ell+1}(y)\ \ \mathscr{T}_{\ell+1}(y)]) \leqslant \ell$. Conversely, (B.1) implies $\widehat{y} \in \widehat{\mathscr{B}}_{\overline{[1,T]}}$ and R palindromic.

B.9. Proof of Proposition 23

The proposition follows from the Sylvester test for coprimness of two scalar polynomials

$$\widehat{c}=\mathrm{GCD}(\widehat{a},\widehat{b}) \text{ and } \deg(\widehat{c}) \geqslant d \iff \operatorname{rank}(\mathscr{R}_{n-d+1}(\widehat{a},\widehat{b})) \leqslant n-d.$$

References

Beck, A., & Ben-Tal, A. (2006). A global solution for the structured total least squares problem with block circulant matrices. *SIAM Journal of Matrix Analysis and Applications*, *27*(1), 238–255.

Berman, A., & Shaked-Monderer, N. (2003). *Completely positive matrices*. Singapore: World Scientific.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Chandrasekaran, S., Gu, M., & Sayed, A. (1998). A stable and efficient algorithm for the indefinite linear least-squares problem. *SIAM Journal of Matrix Analysis and Applications*, *20*(2), 354–362.

De Moor, B. (1993). Structured total least squares and L_2 approximation problems. *Linear Algebra and Its Applications*, *188–189*, 163–207.

De Moor, B. (1994). Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Transactions on Signal Processing*, *42*(11), 3104–3113.

De Moor B. (2005), Dalsy: *Database for the identification of systems*, Department of Electrical Engineering, K.U. Leuven ⟨www.esat.kuleuven.ac.be/sista.daisy/⟩.

Degroat, R., & Dowling, E. (1991). The data least squares problem and channel equalization. *IEEE Transactions on Signal Processing*, *41*, 407–411.

Eckart, G., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Fazel, M. (2002). *Matrix rank minimization with applications*. Ph.D. thesis, Electrical Engineering Department, Stanford University.

Golub, G., & Pereyra, V. (2003). Separable nonlinear least squares: The variable projection method and its applications. *Institute of Physics, Inverse Problems*, *19*, 1–26.

Golub, G., & Van Loan, C. (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, *17*, 883–893.

Graillat, S. (2006). A note on structured pseudospectra. *Journal of Computational and Applied Mathematics*, *191*, 68–76.

Ho, B. L., & Kalman, R. E. (1966). Effective construction of linear state-variable models from input/output functions. *Regelungstechnik*, *14*(12), 545–592.

Kukush, A., Markovsky, I., & Van Huffel, S. (2005). Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. *Journal of Statistical Planning and Inference*, *133*(2), 315–358.

Kung, S. (1978). A new identification method and model reduction algorithm via singular value decomposition. In *Proceedings of the 12th Asilomar conference on circuits, systems, and computers* (pp. 705–714). Pacific Grove, CA.

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Ljung, L. (1999). *System identification: Theory for the user*. Upper Saddle River, NJ: Prentice-Hall.

Manton, J., Mahony, R., & Hua, Y. (2003). The geometry of weighted low-rank approximations. *IEEE Transactions on Signal Processing*, *51*(2), 500–514.

Markovsky, I., & Van Huffel, S. (2005). High-performance numerical algorithms and software for structured total least squares. *Journal of Computational and Applied Mathematics*, *180*(2), 311–331.

Markovsky, I., & Van Huffel, S. (2007). Left vs right representations for solving weighted low rank approximation problems. *Linear Algebra and Its Applications*, *422*, 540–552.

Markovsky, I., Van Huffel, S., & Pintelon, R. (2005). Block-Toeplitz/Hankel structured total least squares. *SIAM Journal of Matrix Analysis and Applications*, *26*(4), 1083–1099.

Markovsky, I., Willems, J. C., De Moor, B. (2006). Comparison of identification algorithms on the database for system identification DAISY. In *Proceedings of the 17th symposium on mathematical theory of networks and systems* (pp. 2858–2869). Kyoto, Japan.

Markovsky, I., Willems, J. C., Rapisarda, P., & De Moor, B. (2005). Algorithms for deterministic balanced subspace identification. *Automatica*, *41*(5), 755–766.

Markovsky, I., Willems, J. C., Van Huffel, S., De Moor, B. (2006). Exact and approximate modeling of linear systems: A behavioral approach. In *Monographs on mathematical modeling and computation*, vol. 11. SIAM.

Markovsky, I., Willems, J. C., Van Huffel, S., De Moor, B., & Pintelon, R. (2005). Application of structured total least squares for system identification and model reduction. *IEEE Transactions on Automatic Control*, *50*(10), 1490–1500.

Pintelon, R., & Schoukens, J. (2001). *System identification: A frequency domain approach*. Piscataway, NJ: IEEE Press.

Polderman, J., & Willems, J. C. (1998). *Introduction to mathematical systems theory*. New York: Springer.

Roorda, B. (1995). Algorithms for global total least squares modelling of finite multivariable time series. *Automatica*, *31*(3), 391–404.

Roorda, B., & Heij, C. (1995). Global total least squares modeling of multivariate time series. *IEEE Transactions on Automatic Control*, *40*(1), 50–63.

Rump, S. (2003). Structured perturbations part I: Normwise distances. *SIAM Journal of Matrix Analysis and Applications*, *25*, 1–30.

Söderström, T. (2007). Errors-in-variables methods in system identification. *Automatica*, *43*, 939–958.

Söderström, T., & Stoica, P. (1989). *System identification*. NJ: Prentice-Hall.

Trefethen, L. N., & Embree, M. (1999). *Spectra and pseudospectra: The behavior of nonnormal matrices and operators*. Princeton, NJ: Princeton University Press.

Van Huffel, S., & Van dewalle, J. (1991). *The total least squares problem: Computational aspects and analysis*. Philadelphia: SIAM.

Van Overschee, P., & De Moor, B. (1996). *Subspace identification for linear systems: theory, implementation, applications*. Boston: Kluwer.

Vanluyten, B., Willems, J. C., De Moor, B. (2005). Model reduction of systems with symmetries. In *Proceedings of the 44th IEEE conference on decision and control* (pp. 826–831). Seville, Spain.

Vanluyten, B., Willems, J. C., & De Moor, B. (2006). Matrix factorization and stochastic state representations. In *Proceedings of the 45th IEEE conference on decision and control* (pp. 4188–4193). San Diego, California.

Wentzell, P., Andrews, D., Hamilton, D., Faber, K., & Kowalski, B. (1997). Maximum likelihood principle component analysis. *Journal of Chemometrics*, *11*, 339–366.

Willems, J. C. (1986, 1987). 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.

Willems, J. C. (1991). Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control*, *36*(3), 259–294.

Ivan Markovsky obtained MS degree in Control and Systems Engineering in 1998 from the Technical University of Sofia, Bulgaria and Ph.D. degree in 2005 from the Katholieke Universiteit Leuven, Belgium. Since January 2007 he is a lecturer at the School of Electronics and Computer Science, University of Southampton, UK. His research work is focused on identification methods in the behavioral setting and errors-in-variables estimation problems. He is a coauthor of the book “Exact and Approximate Modeling of Linear Systems: A Behavioral Approach” (SIAM, Philadelphia, 2006).