

# Software for weighted structured low-rank approximation

Ivan Markovsky and Konstantin Usevich

School of Electronics and Computer Science  
University of Southampton  
{im,ku}@ecs.soton.ac.uk

## Abstract

A software package is presented that computes locally optimal solutions to low-rank approximation problems with the following features:

- *mosaic Hankel structure* constraint on the approximating matrix,
- *weighted 2-norm* approximation criterion,
- *fixed elements* in the approximating matrix,
- *missing elements* in the data matrix, and
- *linear constraints* on an approximating matrix's left kernel basis.

It implements a variable projection type algorithm and allows the user to choose standard local optimization methods for the solution of the parameter optimization problem. For an  $m \times n$  data matrix, with  $n > m$ , the computational complexity of the cost function and derivative evaluation is  $O(m^2n)$ . The package is suitable for applications with  $n \gg m$ . In statistical estimation and data modeling—the main application areas of the package— $n \gg m$  corresponds to modeling of large amount of data by a low-complexity model. Autonomous linear time-invariant system identification and approximate common divisor problems are presented as example applications of the package.

**Keywords:** mosaic Hankel matrix, low-rank approximation, total least squares, system identification, deconvolution, variable projection,

## 1 Introduction

Structured low-rank approximation is defined as low-rank approximation

$$\text{minimize over } \hat{D} \quad \|D - \hat{D}\| \quad \text{subject to} \quad \text{rank}(\hat{D}) \leq r$$

with the additional constraint that the approximating matrix  $\hat{D}$  has the same structure as the data matrix  $D$ . A typical example where a rank deficient structured matrix arises is when a sequence  $p = (p_1, \dots, p_{n_p})$  satisfies a difference equation with lag  $\ell < \lceil n_p/2 \rceil$ , i.e.,

$$R_0 p_t + R_1 p_{t+1} + \dots + R_\ell p_{t+\ell} = 0, \quad \text{for } t = 1, \dots, n_p - \ell. \quad (\text{DE})$$

The system of equations (DE) is linear in the vector of parameters  $R := [R_0 \ R_1 \ \dots \ R_\ell]$ , so that it can be written as  $R \mathcal{H}_{\ell+1, n_p-\ell}(p) = 0$ , where  $\mathcal{H}_{\ell+1, n_p-\ell}(p)$  is a Hankel matrix constructed from  $p$ . This shows that, for  $R \neq 0$ , the fact that  $p$  satisfies a difference equation (DE) is equivalent to rank deficiency of a Hankel matrix  $\mathcal{H}_{\ell+1, n_p-\ell}(p)$ .

Many problems in machine learning, system theory, signal processing, and computer algebra can be posed and solved as structured low-rank approximation problem for different types of structures and different approximation criteria (see Section 5 and [28, 30]). In identification and model reduction of linear time-invariant dynamical systems, the structure is block-Hankel. In the computation of approximate greatest common divisor of two polynomials, the structure is Sylvester. In machine learning, the data matrix is often unstructured but the approximation criterion is a weighted 2-norm (or semi-norm, in the case of missing data).

Despite the academic popularity and numerous applications of the structured low-rank approximation problem, the only efficient publicly available software package for structured low-rank approximation is the one of [35]. The

Feature	Old version	New version
matrix structure	block-Hankel/Toeplitz	matrix $\times$ mosaic-Hankel
cost function	2-norm	weighted 2-norm
exact data	whole blocks	arbitrary elements
missing data	not allowed	arbitrary elements
constraints on the optimization variable	unconstrained	linear constraints
interface	Matlab	Matlab, Octave, R

Table 1: Comparison of the old and new versions of the software.

package presented in this paper is a significantly extended version of the software in [35]. The main extensions are summarized in Table 1 and are described in more details in Section 2.

The paper is organized as follows. Section 3 defines the considered weighted structured low-rank approximation and presents Matlab/Octave and R interfaces for calling the underlying C++ solver. Section 4 gives details about the solution method for solving the resulting parameter optimization problem. Implementation and software design issues are extracted in Appendix A. Section 5 lists applications of the package and describes in more details an application for solving scalar autonomous linear time-invariant identification and approximate common divisor problems. Appendix B defines the structured total least squares problem and explains its relation to structured low-rank approximation. Appendix C lists extra options of the software for choosing the optimization method.

## 2 Main features of the software

### 1. Matrix structure specification

The software package supports matrix structures of form

$$\mathcal{S}(p) := \Phi \mathcal{H}_{\mathbf{m},\mathbf{n}}(p), \quad (\mathcal{S})$$

where  $\Phi$  is a full row rank matrix and  $\mathcal{H}_{\mathbf{m},\mathbf{n}}$  is a *mosaic Hankel* structure [18], i.e., a  $q \times N$  block matrix

$$\mathcal{H}_{[m_1 \dots m_q], [n_1 \dots n_N]}(p) = \begin{bmatrix} \mathcal{H}_{m_1, n_1}(p^{(11)}) & \dots & \mathcal{H}_{m_1, n_N}(p^{(1N)}) \\ \vdots & & \vdots \\ \mathcal{H}_{m_q, n_1}(p^{(q1)}) & \dots & \mathcal{H}_{m_q, n_N}(p^{(qN)}) \end{bmatrix}, \quad (\mathcal{H}_{\mathbf{m},\mathbf{n}})$$

with scalar *Hankel* blocks

$$\mathcal{H}_{m,n}(p) := \begin{bmatrix} p_1 & p_2 & p_3 & \dots & p_n \\ p_2 & p_3 & \ddots & & p_{n+1} \\ p_3 & \ddots & & & \vdots \\ \vdots & & & & \\ p_m & p_{m+1} & \dots & & p_{m+n-1} \end{bmatrix} \in \mathbb{R}^{m \times n}. \quad (\mathcal{H}_{m,n})$$

$(\mathcal{H}_{\mathbf{m},\mathbf{n}})$  is more general than the block-Hankel and “flexible structure specification”, used in the old version of the software (see Section 3 in [35]). In fact, the “flexible structure specification” is equivalent to  $\mathcal{H}_{\mathbf{m},\mathbf{n}}$  with equal  $n_i$ ’s. Mosaic Hankel matrices with blocks of different column dimension allow us to solve, for example, system identification problems with multiple trajectories of different lengths.

The matrix  $\Phi$  further extends the class of mosaic Hankel matrices to (mosaic) Hankel-like matrices. A trivial example is the Toeplitz structure achieved by

$$\Phi = J_m := \begin{bmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{m \times m} \quad \text{and} \quad \mathcal{H}_{\mathbf{m},\mathbf{n}}(p) = \mathcal{H}_{m,n}(p).$$

(Empty spaces in a matrix denote zeros.) A more interesting example is the *Toeplitz-plus-Hankel* structure, achieved by

$$\Phi = \begin{bmatrix} I_m & J_m \end{bmatrix} = \begin{bmatrix} 1 & & & & 1 \\ & \ddots & & & \\ & & 1 & 1 & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \in \mathbb{R}^{m \times 2m} \quad \text{and} \quad \mathcal{H}_{\mathbf{m},\mathbf{n}}(p) = \begin{bmatrix} \mathcal{H}_{m,n}(p^{(1)}) \\ \mathcal{H}_{m,n}(p^{(2)}) \end{bmatrix}. \quad (\mathcal{T} + \mathcal{H})$$

Other examples of using  $\Phi$ , including multipolynomial Sylvester matrices [21], can be found in Section 5.

## 2. Cost function

In the old version of the package, the approximation criterion is the 2-norm (unweighted low-rank approximation). In the new version, the approximation criterion is the weighted 2-norm. Weights are needed for example when the accuracy of the elements of  $p$  vary, see [52]. Weights can be used also to replace the parameter 2-norm with the Frobenius matrix norm of the error matrix  $\mathcal{S}(p - \hat{p})$ . As explained in Note 2, the weighted approximation criterion can be viewed alternatively as a modification of the data matrix structure.

## 3. Exact data

Structure parameters can be fixed to predefined values. (In the old version of the package only subblocks of the structured matrix could be specified as fixed.) This feature allows us to solve problems, such as system identification under exactly known initial conditions, where the upper-left triangles of the Hankel blocks are fixed, and approximate common divisor computation, where the upper-left and lower-right triangles of the Hankel blocks are fixed.

## 4. Missing data<sup>1</sup>

A new feature of the software is the ability to deal with missing data. Due to specific experiment design or data corruption, the data may be incomplete in real-life applications. A well known example from machine learning is prediction of user ratings (recommender system), see, *e.g.*, [4]. A possibility to specify missing values in Hankel low-rank approximation problems allows us to solve system identification problems with incomplete or irregularly sampled data.

## 5. Total least squares and low-rank approximation

In the old version of the package, the optimization variable is the  $X$  matrix in the problem of solving approximately an overdetermined system of linear equations  $AX \approx B$ . This problem may be viewed as a restriction of the low-rank approximation problem to a subclass of matrices. (For more details see Appendix B). To preserve backward compatibility, the new version of the software can also solve the structured total least-squares problems, see Note 7.

## 6. Constraints on the optimization variables

The new version of the package uses as optimization variable a matrix  $R$  whose rows form a basis for the left kernel of the approximation  $\hat{D}$  and allows specification of linear structure constraints on  $R$ . An application where such a constraint is needed is multivariate system identification with fixed observability indexes [16]. Other applications are given in Section 5.

## 7. Interfaces to scientific computing environments

The package provides interface for calling the underlying C++ solver from Matlab, Octave, and R, which are among the most often used computing environments for scientists and engineers.

An advantage of unifying structured and weighted low-rank approximation with missing and fixed values and constraints on the optimization variable is that a single algorithm and a piece of software solves a large variety of problems. Table 2 in Section 5 lists special cases of the generic problem solved by the software package. Each special

---

<sup>1</sup>Currently (June 6, 2012), this feature is implemented only in an experimental Matlab version of the software. It can be simulated in the C++ solver by assigning “small” weights to the missing parameters and replacing the missing values with zeros.

case is motivated by applications, which are considered in the literature and specialized algorithms are developed for their solution. Only a few of the algorithms reported in the literature, however, are implemented in publicly available software.

Fixed values in the approximating matrix can be viewed as an extreme case of weighted low-rank approximation with large weights corresponding to the fixed parameters. Similarly, the missing data case can be viewed as weighted low-rank approximation with small weights corresponding to the missing parameters. Both cases are equivalent to a single singular (zero or infinite weights) weighted low-rank approximation problem and can be solved with the same algorithm.

The generality of the algorithm is not compromised by its computational efficiency. Eliminating the approximation  $\hat{D}$  and exploiting the structure in the resulting nonlinear least squares problem, the cost function and derivative evaluation is done in  $O(m^2n)$  floating point operations [31]. Indeed, the more general algorithm implemented in the package reduces to the algorithm in [35] in the special case treated by the old version of the software.

The software is written in C++ with Matlab/Octave and R interfaces. The C++ code is complemented by an experimental implementation written in Matlab. The latter supports general linearly structured matrix as and experimental features but is inefficient and can be used only for small size (say,  $m \leq n < 100$ ) problems. In Matlab, the C++ solver can be called directly via the mex-file `slra_mex` or via a wrapper function, which has the same interface as the mex file but can optionally call the experimental version `slra_ext`. The package is hosted at the following address:

<http://github.com/slra/slra>

### 3 Problem formulation

We denote missing data values by the symbol NaN (“not a number”) and define the extended set of real numbers  $\mathbb{R}_e$  as the union of the set of the real numbers  $\mathbb{R}$  and the symbol NaN:

$$\mathbb{R}_e := \mathbb{R} \cup \text{NaN}.$$

The considered structured low-rank approximation problem is defined as follows.

**Problem 1.** Given:

- structure specification  $\mathcal{S}$ ,
- vector of structure parameters  $p \in \mathbb{R}_e^{n_p}$ ,
- nonnegative vector  $w \in (\mathbb{R}_+ \cup \{\infty\})^{n_p}$ , defining a weighted (semi-)norm

$$\|p\|_w^2 := \sum_{\{i \mid p_i \neq \text{NaN}\}} w_i p_i^2, \text{ and}$$

- desired rank  $r$ ,

find a structure parameter vector  $\hat{p}$ , such that the corresponding matrix  $\mathcal{S}(\hat{p})$  has rank at most  $r$ , and is as close as possible to  $p$  in the sense of the weighted semi-norm  $\|\cdot\|_w$ , i.e.,

$$\text{minimize over } \hat{p} \in \mathbb{R}_e^{n_p} \quad \|p - \hat{p}\|_w^2 \quad \text{subject to} \quad \text{rank}(\mathcal{S}(\hat{p})) \leq r. \quad (\text{SLRA})$$

Without loss of generality, it is assumed that  $m \leq n$ . An infinite weight  $w_i = \infty$  imposes the equality constraint  $\hat{p}_i = p_i$  on the optimization problem (SLRA).

Problem (SLRA) is in general nonconvex. In the software package, it is solved numerically by local optimization methods. The Matlab wrapper function for low-rank approximation is defined as follows:

```
<slra function definition in Matlab/octave>≡
function [ph, info] = slra(p, s, r, varargin)
```

The output argument `ph` of the `slra` function is a locally optimal solution  $\hat{p}$  of (SLRA) and `info` is a structure containing additional information about the computed solution:

- `info.Rh` is an  $(m-r) \times r$  full row rank matrix  $\hat{R}$ , such that  $\hat{R}\mathcal{S}(\hat{p}) = 0$  (low-rank certificate),
- `info.fmin` is the cost function value  $\|p - \hat{p}\|_w^2$  at the computed solution,
- `info.iter` is the number of iterations performed by the optimization solver, and
- `info.time` is the execution time.

Next we describe the required input parameters `p`, `s`, and `r`. Appendix C describes the optional parameter `opt`.

### Matrix structure specification

The structure ( $\mathcal{S}$ ) is specified by the two vectors

$$\mathbf{m} := [m_1 \quad \cdots \quad m_q]^\top \in \mathbb{N}^q \quad \text{and} \quad \mathbf{n} := [n_1 \quad \cdots \quad n_N]^\top \in \mathbb{N}^N \quad (\mathbf{m}, \mathbf{n})$$

and the matrix  $\Phi \in \mathbb{R}^{m \times m'}$ , where

$$m' := \sum_{i=1}^q m_i.$$

The compulsory input arguments `p`, `s`, and `r` of the `slra` function correspond to the vector of parameters  $p$ , the problem structure, and the rank  $r$ , respectively. `s` should be a structure with a field `m` and, optionally, a field `n`, containing the vectors  $(\mathbf{m}, \mathbf{n})$ . If `s.n` is skipped, by default  $N = 1$  and

$$\mathbf{n} = n = \frac{n_p - \sum_{i=1}^q m_i}{q} + 1.$$

Optionally, `s` can have a field `phi`, containing the matrix  $\Phi \in \mathbb{R}^{m \times m'}$ . The default value for  $\Phi$  is the identity matrix of size  $m = m'$ .

The vector  $p$  is composed of the structure parameter vectors of all scalar Hankel blocks in  $(\mathcal{H}_{\mathbf{m}, \mathbf{n}})$ , ordered first top to bottom and then left to right, *i.e.*,

$$p = (p^{(11)}, \dots, p^{(q1)}, \dots, p^{(1N)}, \dots, p^{(qN)}).$$

### Weight specification, fixed and missing parameter values

The weight vector  $w$  is passed to the `slra` function by the field `w` of `s`. If  $w$  is not specified, its default value is the vector of all ones, corresponding to unweighted low-rank approximation. The parameter `s.w` can be

- $n_p$ -dimensional nonnegative vector, in which case  $w = \mathbf{s}.w$ ,
- $q \times N$  nonnegative matrix, in which case the weights corresponding to the structure parameters of the  $i, j$ th Hankel block in  $(\mathcal{H}_{\mathbf{m}, \mathbf{n}})$  are all equal to  $w_{ij}$ , or
- $q$ -dimensional nonnegative vector, in which case the weights corresponding to the structure parameters of the  $i, j$ th Hankel block in  $(\mathcal{H}_{\mathbf{m}, \mathbf{n}})$  are all equal to  $w_i$ , for all  $j$ .

The second and third options of specifying `s.w` evoke a more efficient computational method but are supported only by the C++ solver.

Setting  $w_i$  to  $\infty$  has the effect of specifying the parameter  $p_i$  as exact, *i.e.*, the approximation problem (SLRA) is solved with the additional constraint that  $\hat{p}_i$  is equal to the given parameter value  $p_i$  (and the weighted norm  $\|\cdot\|_w$  is evaluated over the parameters with finite weights only). Setting  $w_i$  to 0 has the effect of ignoring the value of the parameter  $p_i$  in the solution of the approximation problem. Alternatively, ignored parameter values may be specified by the symbol `NaN`. From a practical point of view, ignored structure parameter values are “missing data”.

*Note 2* (Incorporating the weights  $w$  into the matrix structure  $\mathcal{S}$ ). A weighted structured problem (SLRA) can be solved as an equivalent unweighted problem (SLRA), *i.e.*, with weights  $w_i = 1$  for all  $i$ , and a modified structure specification  $\mathcal{S}' : \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{m \times n}$ , defined by

$$\mathcal{S}'((p_1, \dots, p_{n_p})) = \mathcal{S}((w_1 p_1, \dots, w_{n_p} p_{n_p})).$$

This fact shows that the weighted norm specification can be viewed equivalently as a modification of the *data* structure.

*Note 3* (Conversion from the structure specification in [35]). In order to facilitate the transition from the old version of the software to the new one, we provide in the package a function `s_old2s` converting the old style structure specification (see [35, page 314]) to the new one. Block structured matrices consisting of block-Hankel and unstructured blocks are specified solely by the  $(\mathbf{m}, \mathbf{n})$  parameters. Block-Toeplitz matrices are specified as block-Hankel matrices with suitable row permutations. In this case, in addition to the parameters  $(\mathbf{m}, \mathbf{n})$ , the transformation matrix  $\Phi$  is used. Finally, fixed blocks are specified by infinite elements in the weight vector  $w$ . In general, the transition from the old structure specification format to the new one involves the  $(\mathbf{m}, \mathbf{n})$ ,  $\Phi$ , and  $w$  parameters.

*Note 4* (R interface). Calling the C++ solver from R is done by an R function `slra` with the same parameters as in the Matlab mex function, except that in R lists are used instead of structures, and the list elements are accessed by `<list>$<element>` instead of by `<structure>.<element>`. The computed result is returned in a list with fields `ph` and `info`, which contain, respectively,  $\hat{p}$  and the `info` variable defined above. Optional logical parameters `compute.ph` and `compute.Rh` can be used to disable the computation of  $\hat{p}$  and  $\hat{R}$ . By default,  $\hat{p}$  is not computed. Examples of using the R interface can be found in the directory `test_r` of the package distribution.

## 4 Solution method

This section outlines the method used in the package for solving the structured low-rank approximation problem (SLRA). Let  $d$  be the rank reduction

$$d := m - r.$$

The rank constraint is represented in a kernel form

$$\text{rank}(\mathcal{S}(\hat{p})) \leq r \quad \Longleftrightarrow \quad \text{there is full row rank } R \in \mathbb{R}^{d \times m}, \text{ such that } R\mathcal{S}(\hat{p}) = 0, \quad (\text{KER})$$

*i.e.*, the left kernel of  $\mathcal{S}(\hat{p})$  is parameterized by the  $d = m - r$  linearly independent rows of the matrix  $R$ . Then (SLRA) is equivalent to the following double minimization problem

$$\text{minimize} \quad \text{over full row rank } R \in \mathbb{R}^{d \times m} \quad f(R), \quad (\text{SLRA}_R)$$

where

$$f(R) := \left( \min_{\hat{p} \in \mathbb{R}^{n_p}} \|p - \hat{p}\|_w^2 \quad \text{subject to} \quad R\mathcal{S}(\hat{p}) = 0 \right). \quad (f(R))$$

(SLRA<sub>R</sub>) is referred to as the outer minimization and  $(f(R))$  as the inner minimization. The inner minimization problem is a (generalized) linear least-norm problem and can be solved analytically for given  $R$ , resulting in a closed form expression of the cost function  $f$  in the outer minimization problem. The derivative of  $f$  with respect to the elements of  $R$  can also be computed analytically. Fast methods for evaluation of  $f$  and its derivatives are presented in Section 4.2.

In data modeling—the main application area of the software— $R$  can be interpreted as a model parameter. For example, in system identification,  $R$  is related to a difference equation (DE) representation of the system. Typically, a large amount of data is fitted by a simple model, which implies that the dimension of  $R$  is small compared to the dimension of  $\hat{p}$ . Problem  $(f(R))$  is bilinear in the variables  $R$  and  $\hat{p}$ , which makes possible the elimination of  $\hat{p}$  for a fixed  $R$ . Many estimation problems have similar bilinear structure and their solution methods are based on elimination of the large dimensional optimization variable.

The elimination of  $\hat{p}$  in the analytic solution of  $(f(R))$  is similar to the projection step in the variable projections method [12] for solving separable nonlinear least squares problems. Note, however, that the standard variable projections method is invented as a solution method for *unconstrained* optimization problems, where the cost function



is explicitly given. The structured low-rank approximation problem (SLRA) in contrast is a constraint optimization problem and the parameterization (KER) is an implicit function of the optimization variables. As discussed next, (SLRA) is an optimization problem over a Grassmann manifold, which makes it rather different from the standard setup, for the variable projections method.

*Note 5* (Feasibility of the minimization problem over  $\hat{p}$ ). In the analytic solution of  $(f(R))$ , it is assumed that the problem is feasible for any  $R$  in the search space. A necessary condition for this to hold is that the number of structure parameters is sufficiently large:

$$n_p > dn. \quad (\text{C})$$

Condition (C) imposes restrictions on the class of problems that can be solved by the approach implemented in the package. For example, in scalar Hankel low-rank approximation problems the rank reduction  $d$  can be at most one. Interestingly, in applications of problem (SLRA) to approximate realization, system identification, and approximate common divisor computation constraint (C) can always be satisfied.

#### 4.1 Parametrization of the search space and constraints

Problem (SLRA<sub>R</sub>) is an optimization problem on a *Grassmann manifold*  $\text{Gr}(d, m)$  (set of  $d$ -dimensional subspaces of  $\mathbb{R}^m$ ). Indeed,  $f(R_1) = f(R_2)$  if the rows of  $R_1$  and  $R_2$  span the same subspace. In the package, we map the optimization on  $\text{Gr}(d, m)$  to an optimization problem on an Euclidean space, see [48]. The problem is turned into optimization of  $f$  over the set of full row rank matrices  $R \in \mathbb{R}^{d \times m}$  that represent all (or a generic part) of the  $d$ -dimensional subspaces.

In some applications, it is necessary to impose linear constraints on  $R$ , which can also be incorporated into the parametrization of the search space.

- The experimental Matlab solver uses a general linear constraint on  $R$

$$R = \mathcal{R}'(\theta) := \text{vec}_d^{-1}(\theta\Psi), \quad \text{where } \theta \in \mathbb{R}^{n_p}, \quad (\theta \mapsto R)$$

defined by a matrix  $\Psi \in \mathbb{R}^{n_\theta \times dm}$ . (Here  $\text{vec}(\cdot)$  is the column-wise vectorization operator and  $\text{vec}^{-1}(\cdot)$  is its inverse.) The rank constraint on  $R$  is imposed by

$$RR^\top = I_d. \quad (\text{f.r.r. } R)$$

- The C++ solver uses a matrix-product constraint on the matrix  $R$

$$R = \mathcal{R}(\Theta) := \Theta\Psi, \quad \text{where } \Theta \in \mathbb{R}^{d \times m''}, \quad (\Theta \mapsto R)$$

defined by a matrix  $\Psi \in \mathbb{R}^{m'' \times m}$ . The rank constraint on  $R$  is imposed by

$$\Theta = \begin{bmatrix} X & -I_d \end{bmatrix}, \quad \text{for some } X \in \mathbb{R}^{d \times (m''-d)}. \quad (\Theta \leftrightarrow X)$$

The  $\Psi$  matrix is passed to the `slra` function by an optimization option `opt.psi`. The default value of  $\Psi$  is  $I_m$  for the C++ version and  $I_{dm}$  for the Matlab version. The  $\Psi$  matrices— $\Psi_C$  for the C++ versions and  $\Psi_M$  for the Matlab versions—are generally different. They coincide if and only if the rank reduction is one, *i.e.*,  $d = 1$ . The constraint  $(\theta \mapsto R)$  is more general and includes the constraint  $(\Theta \mapsto R)$  by choosing

$$\Psi_M = \Psi_C \otimes I_d \quad \text{and} \quad \theta = \text{vec}(\Theta).$$

The rank constraint (f.r.r.  $R$ ) parametrizes  $\text{Gr}(d, m)$ , and turns the structured low-rank approximation problem into an optimization problem on  $\mathbb{R}^{md}$  with a quadratic constraint. The rank constraint  $(\Theta \leftrightarrow X)$  parametrizes a generic part of  $\text{Gr}(d, m)$ . Its main advantage is that the structured low-rank approximation problem is turned into an unconstrained optimization problem on  $\mathbb{R}^{(m''-d)d}$ .

*Note 6* (On parametrization of the whole  $\text{Gr}(d, m)$  in the C++ version). The whole  $\text{Gr}(d, m)$  can be covered by taking different  $\Psi$  (see [48]) or changing  $\Psi$  during the optimization, which is done in the optimization algorithms of [27]. Currently, in the C++ version,  $\Psi$  is a fixed parameter.

*Note 7* (Solution of structured total least squares problems). Setting  $\Psi = I_d$  in the C++ version is equivalent to solving a structured total least squares problem (see Appendix B). In this case,  $\hat{R} = \begin{bmatrix} X & -I_d \end{bmatrix}$ .

## 4.2 Fast cost function evaluation

In this section we present the algorithm for the cost function evaluation in the case of positive weights (no missing values), which is implemented in C++ solver. The cost function evaluation for missing data is described in [31].

The cost function can be represented in the form

$$f(R) := \text{vec}^\top(R\mathcal{S}(p)) (\Gamma(R))^{-1} \text{vec}(R\mathcal{S}(p)),$$

where  $\Gamma(R) \in \mathbb{R}^{dn \times dn}$  is a nonnegative definite matrix, depending on  $R$ . For the evaluation of  $f$  (and its derivatives) one needs to solve the system of linear equations

$$\Gamma(R)u = v(R), \quad \text{where } v := \text{vec}(R\mathcal{S}(p)).$$

In the package, we use the Cholesky factorization of  $\Gamma$ , which is known to have good numerical properties [14]. The efficiency of the package is based on the fact that  $\Gamma$  is block banded with bandwidth  $s = \min(m_1, \dots, m_N)$  and, in the case of uniform weights,  $\Gamma$  is in addition block-diagonal with block-Toeplitz elements:

$$\Gamma(R) := \begin{bmatrix} \Gamma_{1,1} & \cdots & \Gamma_{1,s} & & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ \Gamma_{s,1} & \ddots & \ddots & \ddots & \Gamma_{n-s+1,n} \\ & \ddots & \ddots & \ddots & \vdots \\ 0 & & \Gamma_{n,n-s+1} & \cdots & \Gamma_{n,n} \end{bmatrix}, \quad \text{where } \Gamma_{i,j} = RW_{i,j}R^\top \in \mathbb{R}^{d \times d},$$

and the matrices  $W_{i,j} \in \mathbb{R}^{m \times m}$  depend only on the structure specification and the weights. For  $j \geq i$  and structure of form  $\mathcal{H}_{\mathbf{m},n}$ , where  $\mathbf{m} = [m_1 \dots m_q]$ , the matrices  $W_{i,j}$  can be represented as

$$W_{i,j} = \begin{cases} (U_{\mathbf{m}})^{j-i} & w = \mathbf{1} \text{ (uniform weights) and} \\ \text{diag}(\gamma_{i,j}) (U_{\mathbf{m}})^{j-i} & \text{otherwise,} \end{cases}$$

where  $\gamma_{i,j} \in \mathbb{R}^m$  is a subvector of  $\text{col}(w_1^{-1}, \dots, w_{n_p}^{-1})$ , and

$$U_{\mathbf{m}} := \text{diag}(U_{m_1}, \dots, U_{m_q}), \quad \text{where } U_m := \begin{bmatrix} 0 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 1 \\ 0 & \dots & \dots & 0 \end{bmatrix}.$$

For a general structure of the form  $\Phi\mathcal{H}_{\mathbf{m},\mathbf{n}}$ , where  $\mathbf{m}$  and  $\mathbf{n}$  are defined in  $(\mathbf{m}, \mathbf{n})$ , the  $\Gamma$  matrix is block-diagonal

$$\Gamma(R) = \text{diag}(\Gamma^{(1)}(R\Phi), \dots, \Gamma^{(N)}(R\Phi)),$$

with diagonal block  $\Gamma^{(k)}(R)$  being the  $\Gamma$  matrix for the structure  $\mathcal{H}_{\mathbf{m},n_k}$  and the corresponding subvector of  $w$ .

In summary, the cost function can be computed by the following algorithm:

1. Compute  $v(R) = \text{vec}(R\mathcal{S}(p))$ .
2. Compute  $\Gamma(R)$ , using  $W_{i,j}$ .
3. Compute the Cholesky factorization  $\Gamma(R) = C^\top C$  ( $\Gamma$  is an  $(2sd)$ -banded matrix).
4. Compute  $F(R) = C^{-\top} v(R) \in \mathbb{R}^{nd}$  by solving  $(sd)$ -banded system of linear equations.
5. Set  $f(R) := \|F(R)\|_2^2$ .

For general weights, the algorithm has complexity  $O(smn)$ . For uniform weights, the  $\Gamma$  matrix is block-diagonal with block-Toeplitz blocks and the algorithm has theoretical complexity  $O(mn)$ , assuming that an efficient method (e.g., the Schur-type algorithms [20]) is using for the Cholesky factorization of the block-Toeplitz matrices.



### 4.3 Local optimization methods

Due to the elimination of  $\hat{p}$ , for  $n \gg m$ , the computation time of the search direction by the optimization method is negligible in comparison with the computation time for the cost function and derivative evaluations. Therefore, we use standard local optimization methods to minimize the cost function  $f$ .

The C++ solver uses local optimization methods from the GSL library (see Appendix C). One option is to use the Nead-Melder method for minimization without derivatives, which is slow but robust. A faster alternative is to use the BFGS quasi-Newton method. The cost function gradient is computed analytically by

$$df(R, H) = 2y^\top dv(R, H) - y^\top d\Gamma(R, H)y, \quad \text{where } y := \Gamma^{-1}v(R).$$

It can be shown that the gradient evaluation has computational complexity of the same order as the one of the cost function evaluation.

The default optimization method for the C++ solver is the Levenberg-Marquardt algorithm [39, 8], which is typically faster and more robust in practice than the quasi-Newton methods. The Jacobian of  $C^{-\top}v$ , where  $C$  is the Cholesky factor of  $\Gamma$  is replaced by the pseudo-Jacobian [17], which computation is of the order of  $O(smn)$ .

The package provides an interface to the cost function and derivatives evaluation, so that any optimization method (implemented in or callable from C/C++) can be used. More details about the C++ interface can be found in Appendix A. The experimental Matlab solver uses the function `fmincon` for constrained nonlinear minimization from the Optimization Toolbox of Matlab. In this case, the derivatives are approximated numerically.

All local optimization methods require an initial approximation. By default, unstructured rank- $r$  approximation  $\hat{D}_{\text{lra}}$  of the data matrix  $D = \mathcal{S}(p)$  is used, *i.e.*, the initial value for the parameter  $R$  is a full row rank matrix  $R_{\text{lra}} \in \mathbb{R}^{d \times m}$ , such that  $R_{\text{lra}} \hat{D}_{\text{lra}} = 0$ . The default value can be overwritten by an optional argument `opt.Rini`.

In order to obtain the  $\theta$  or  $\Theta$  parameter from  $R$ , the equation  $\mathcal{R}(\theta) \approx R_{\text{lra}}$  is solved in the least-squares sense. For the C++ solver, the obtained matrix  $\Theta$  is then converted to  $X$  by

$$X := -P^{-1}Q, \quad \text{where } \Theta =: [Q \ P], \quad Q \in \mathbb{R}^{d \times (m''-d)}, \quad \text{and } P \in \mathbb{R}^{d \times d}.$$

The result of the `slra` function—a locally optimal value of the parameter  $R$ —is returned in the variable `info.R`. The corresponding locally optimal approximation  $\hat{p}$  is computed by solving the inner minimization problem ( $f(R)$ ). The solution is given by

$$\hat{p}(R) = p + G(R)(\Gamma(R))^{-1}v(R),$$

where  $G(R)$  is a linear function in the elements of  $R$ , defined by

$$\text{vec}(R\mathcal{S}(p)) = G(R)p.$$

## 5 Applications

A list of examples of low-rank approximation problems with different matrix structures  $\Phi\mathcal{H}_{\mathbf{m},\mathbf{n}}$ , approximation criteria  $\|\cdot\|_w$ , and constraints on a basis of the approximating matrix's left kernel is given in Table 2. Each example is motivated by applications. (See the references cited in the table.) In Section 5.1, we describe the application to autonomous system identification (example 3) in the case of scalar (single output) linear time-invariant dynamical systems. Variations of the problem for identification with fixed poles (example 6), poles on the unit circle (example 4), and missing data (examples 10 and 17) are presented in Sections 5.2, 5.3 and 5.4. Section 5.5 demonstrates the efficiency of the package on data sets from the database for system identification DAISY [6]. Section 5.6 shows numerical examples of computing approximate common divisors of two polynomials (example 14), using Sylvester structured low-rank approximation and Section 5.7 generalizes the results for more than two polynomials (example 15).

### 5.1 Autonomous system identification

In system theory and signal processing the difference equation (DE) defines an autonomous linear time-invariant dynamical system. The problem of finding a model, which is a dynamical system, from observed data of the system

#	example	application(s)	reference
1	unstructured uniform weights	factor analysis subspace methods latent semantic analysis	[15] [37, Ch. 7] [9]
2	element-wise weights	chemometrics	[52]
3	scalar Hankel	model reduction autonomous SYSID linear prediction shape from moments	[3] [37, Ch. 11] [24] [41, 11]
4	Hankel with structured kernel (palindromic)	spectral estimation	[47]
5	Toeplitz + Hankel	isospectral flow	[43]
6	Hankel with structured kernel (fixed poles)	SYSID with some predefined poles	[29]
7	$q \times 1$ block Hankel	multivariable SYSID	[38]
8	Hankel with structured kernel (fixed obsrv. indexes)	SYSID with fixed obsrv. indexes	[16]
9	$q \times N$ block Hankel	SYSID from multiple time series ( $T_i = T$ )	[38]
10	Hankel blocks next to each other	SYSID from multiple time series	[42]
11	Hankel with fixed variables	output error SYSID	[38]
12	Unstructured beneath Hankel block	deconvolution	[28]
13	Hankel with first $\ell$ elements fixed	SYSID with fixed initial conditions	[46]
14	Sylvester	approximate GCD of two polynomials	[22]
15	generalized Sylvester	approximate GCD of $N$ polynomials	[21]
16	unstructured with missing elements	recommender systems	[45]
17	Hankel with missing elements	SYSID with missing data	[42]
18	missing and fixed elements	matrix completion	[5]

Table 2: Summary of examples. (SYSID — system identification, GCD — greatest common divisor)

is called a system identification problem [26, 46]. The particular identification problem considered in this section is defined as follows: Given a time series

$$y_d = (y_d(1), \dots, y_d(T)) \in \mathbb{R}^T$$

and a natural number  $\ell$ ,

$$\begin{aligned} & \text{minimize} && \text{over } \hat{y} \in \mathbb{R}^T \text{ and } [R_0 \ R_1 \ \dots \ R_\ell] \neq 0 && \|y_d - \hat{y}\|_2^2 \\ & \text{subject to} && R_0 \hat{y}(t) + R_1 \hat{y}(t+1) + \dots + R_\ell \hat{y}(t+\ell) = 0, && \text{for } t = 1, \dots, T - \ell. \end{aligned} \quad (\text{SYSID})$$

The constraints of (SYSID) are equivalent to

$$\text{rank}(\mathcal{H}_{\ell+1, T-\ell}(\hat{y})) \leq \ell.$$

Therefore, (SYSID) is equivalent to the scalar Hankel low-rank approximation problem

$$\begin{aligned} & \text{minimize} && \text{over } \hat{y} \in \mathbb{R}^T && \|\mathbf{y}_d - \hat{y}\|_2^2 \\ & \text{subject to} && \text{rank}(\mathcal{H}_{\ell+1, T-\ell}(\hat{y})) \leq \ell, \end{aligned} \quad (\text{SHLRA})$$

which is a special case of (SLRA) with

$$\mathbf{m} = \ell + 1 \quad \text{and} \quad \mathbf{n} = T - \ell.$$

```
<slra arguments for scalar Hankel problem>≡
s.m = ell + 1; s.n = T - ell; r = ell;
```

### Numerical example

The sequence  $\bar{y}$ , shown in Figure 1, left, is a sum of 8 complex exponentials. It follows that it is an exact trajectory of an autonomous linear time-invariant system with lag  $\ell = 8$ .

```
<example autonomous SYSID>≡
clear all; randn('seed', 0); load aut_sys_traj; T = length(y0); ell = 8;
plot(y0(1:T), 'r-'), ax = axis; axis([0 T ax(3:4)]), print_fig('slra-f1')
```

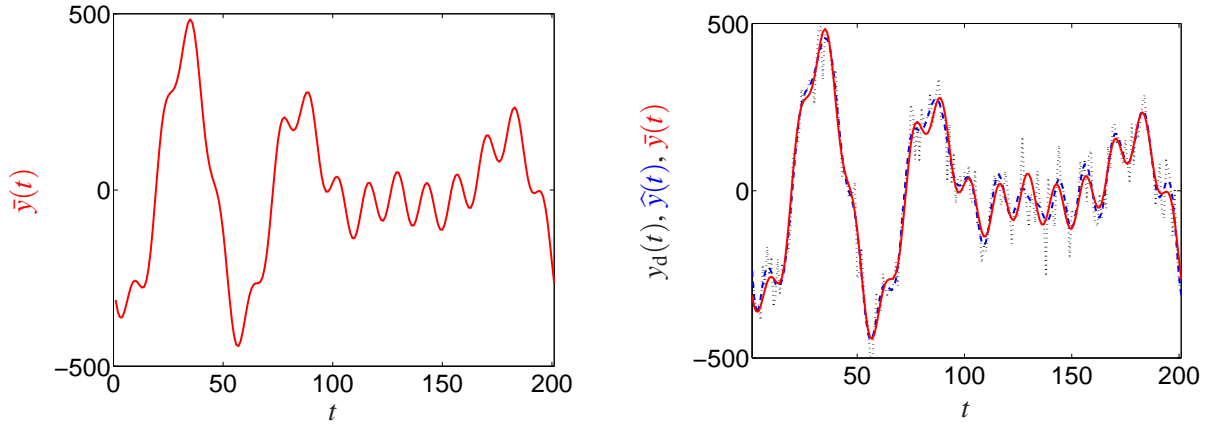


Figure 1: Left: true trajectory. Right: noisy trajectory (dotted black) and optimal approximation (dashed blue).

Indeed the left kernel of the Hankel matrix  $\mathcal{H}_{\ell+1, T-\ell}(\bar{y})$  has dimension one.

```
<example autonomous SYSID>+≡
size(null(blkhank(y0(1:T), ell + 1)'))', 1)
```

Moreover, a basis vector  $\bar{R}$  for the left kernel of  $\mathcal{H}_{\ell+1, T-\ell}(\bar{y})$  gives a difference equation representation with lag  $\ell$  (DE) of the exact model for  $\bar{y}$ .

```
<example autonomous SYSID>+≡
y2R = @(y, ell) null(blkhank(y, ell + 1)'))'; R0 = y2R(y0, ell);
```

Approximation is needed when an exact model does not exist. In statistical estimation problems, this is due to the data being corrupted by noise

$$\mathbf{y}_d = \bar{\mathbf{y}} + \tilde{\mathbf{y}},$$

where  $\bar{y}$  is the true component and  $\tilde{y}$  is the noise.

In the simulation example, the true system belongs to the model class (autonomous systems with lag 8) and the noise  $\tilde{y}$  is generated as a zero mean white Gaussian process.

```
<example autonomous SYSID>+≡
y = y0 + 0.3 * std(y0) * randn(length(y0), 1);
```

Under this assumption, the optimal approximation problem (SYSID) defines a maximum likelihood estimator [28].

```
<example autonomous SYSID>+≡
<slra arguments for scalar Hankel problem>
<define y2R_appr>, opt.Rini = y2R_appr(y(1:T), ell);
[yh, info] = slra(y(1:T), s, r, opt);
```

Figure 1, right, shows the true data, the noisy data, and the optimal fit obtained by the `slra` function.

```
<example autonomous SYSID>+≡
plot(yh, 'b-'), hold on, plot(y(1:T), 'k:'), plot(y0(1:T), 'r-')
axis([1 T ax(3:4)]), print_fig('slra-f2')
```

## 5.2 Autonomous system identification with fixed poles

An  $\ell + 1$  dimensional vector  $R$  corresponds to a degree  $\ell$  polynomial

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_\ell \end{bmatrix} \leftrightarrow R_0 + R_1 z + \cdots + R_\ell z^\ell. \quad (R(z))$$

Vice verse, a polynomial of degree  $\ell$  corresponds to the  $\ell + 1$  dimensional vector formed of its coefficients in, say, increasing order. With some abuse of notation, we denote by  $R$  both the vector and the polynomial.

Consider the scalar Hankel low-rank approximation problem (SHLRA) and let  $R$  be a basis for the left kernel of the approximating matrix  $\mathcal{H}_{\ell+1, T-\ell}(\hat{y})$ . Denote by  $z_1, \dots, z_\ell$  the zeros of  $R$ . Since  $R$  is real, the zeros appear in complex conjugate pairs. Otherwise,  $z_1, \dots, z_\ell$  are unconstrained by the approximation problem (SHLRA). In this section, we impose a constraint that some zeros of  $R$  are fixed at predefined locations  $z_{f,1}, \dots, z_{f,\ell_f}$  in the complex plane (observing the complex conjugate symmetry) and the others are unconstrained. The fixed zero locations define a real polynomial

$$R_f(z) := \prod_{i=1}^{\ell_f} (z - z_{f,i})$$

and the free zeros define a polynomial  $\theta(z)$  of degree  $\ell - \ell_f$ . With this notation, the fixed zeros constraint is

$$R(z) = \theta(z) R_f(z),$$

or in matrix notation

$$R = \theta \underbrace{\begin{bmatrix} R_{f,0} & R_{f,1} & \cdots & R_{f,\ell_f} \\ & R_{f,0} & R_{f,1} & \cdots & R_{f,\ell_f} \\ & & \ddots & \ddots & \ddots \\ & & & R_{f,0} & R_{f,1} & \cdots & R_{f,\ell_f} \end{bmatrix}}_{\Psi}.$$

The resulting scalar Hankel low-rank approximation problem with fixed poles is of the form  $(\Theta \mapsto R)$ , with the  $\Psi$  matrix defined above.

```
<opt.psi for fixed poles>≡
opt_f.psi = sylvester(poly(zf), 1, ell - length(zf) + 1);
```

### Numerical example

Data centering, *i.e.*, subtracting the mean of the data vector  $y_d$ , is a standard preprocessing procedure in linear system identification. A linear model is identified from the centered data and in a postprocessing step the mean is added back to the model in order to obtain an *affine model* for the original data. The preprocessing procedure splits the affine system identification problem into two steps: data centering and linear system identification. A new problem is reduced to an already solved problem, however, the procedure is a suboptimal method for affine system identification. This fact is illustrated on the data from Section 5.1.

```
<example autonomous SYSID>+≡
yc = y(1:T) - mean(y(1:T)); opt_c.Rini = fliplr(poly(eig(h2ss(yc, ell))));
[yhc, info_c] = slra(yc, s, r, opt_c); info_c.fmin
```

The approximation error achieved with the two stage procedure is `info_c.fmin = 3.7169e+05`.

Observing that an affine autonomous model is a linear autonomous model with a fixed pole at one, in the discrete-time case, or zero in the continuous-time case, we obtain an optimal method for affine autonomous system identification as a special case of Hankel low-rank approximation with fixed poles. Applied on the data

```
<example autonomous SYSID>+≡
    ell = ell + 1; <slra arguments for scalar Hankel problem>
    zf = 1; <opt.psi for fixed poles>
    opt_f.Rini = opt_c.Rini * opt_f.psi;
    [yh_f, info_f] = slra(y(1:T), s, ell, opt_f); info_f.fmin
the method achieves approximation error info_c.fmin = 3.6652e+05.
```

### 5.3 Harmonic retrieval

In the autonomous system identification problem (SYSID) the poles of the identified system (*i.e.*, the roots of the polynomial  $R(z)$ ) are constrained only to have complex conjugate symmetry. In this section, we consider a variation of problem (SYSID) when the poles are constrained to lie on the unit circle. This additional constraint ensures that the approximation is of the form of a sum of sinusoidal functions. The problem is referred to as harmonic retrieval and line spectral estimation [47].

A necessary condition for the poles of the system to lie on the unit circle is that  $R$  is palindromic, *i.e.*,

$$\text{there is } \theta \in \mathbb{R}^{1 \times n_\theta}, \quad n_\theta := \left\lceil \frac{\ell+1}{2} \right\rceil, \quad \text{such that } R = \theta \Psi,$$

where  $\lceil \cdot \rceil$  is the operator rounding up to the nearest integer and

$$\Psi = \begin{bmatrix} 1 & & & & 1 \\ & \ddots & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}, \text{ if } m \text{ is odd, or } \Psi = \begin{bmatrix} 1 & & & & 1 \\ & \ddots & & & \\ & & \ddots & & \\ & & & 1 & 1 \\ & & & & \ddots \end{bmatrix}, \text{ if } m \text{ is even.}$$

This is a special case of the constraint ( $\theta \mapsto R$ ) and therefore can be specified in the solution of the identification problem (SHLRA).

```
<opt.psi for palindromic R>≡
    nt = round((ell + 1) / 2); I = eye(nt);
    opt_p.psi = [I flipud(I(:, (1 + mod(ell + 1, 2)):end))];
```

Note that the constraint that  $R$  is palindromic is only a necessary condition for the roots of  $R$  to be on the unit circle. In the following numerical example we verify on the data of Section 5.1 that the model identified by the `slra` function has poles on the unit circle.

#### Numerical example

```
<example autonomous SYSID>+≡
    ell = 8; <slra arguments for scalar Hankel problem>
    <opt.psi for palindromic R>
    opt_p.Rini = fliplr(poly(eig(approx(y(1:T), ell))));
    [yh_p, info_p] = slra(y(1:T), s, ell, opt_p);
    abs(roots(info_p.Rh))
```

### 5.4 Autonomous system identification with missing data

A block of  $\ell$  or more missing values splits the data into  $N = 2$  independent datasets. With  $N - 1$  such blocks of missing data the problem is equivalent to identification from  $N$  datasets. Note, however, that in general the datasets have different number of samples  $T_1, \dots, T_N$ . System identification from multiple datasets with different lengths leads to a mosaic Hankel low-rank approximation problem with

$$\mathbf{m} = \ell \quad \text{and} \quad \mathbf{n} = [T_1 - \ell \quad \dots \quad T_N - \ell].$$

In the more general case of arbitrary distributed missing values, the identification problem is posed as a weighted scalar Hankel low-rank approximation problem. Let  $\mathcal{I}$  be the indexes of the missing samples and  $\bar{\mathcal{I}}$  be the indexes of the given samples. The appropriate weight vector is:

$$w \in \mathbb{R}^T, \quad \text{with } w(\mathcal{I}) = 0 \quad \text{and} \quad w(\bar{\mathcal{I}}) = 1.$$

In the current version of the C++ solver, zero weights should be approximated by small values.

```
<s.w for missing values>≡
s.w = ones(T, 1); s.w(Im) = 1e-5;
```

### Numerical example

First, we select  $N$  continuous parts of the signal and identify a model from this data.

```
<example autonomous SYSID>+≡
ell = 8; T = [20 30 40]; N = length(T); ind = [1 50 100];
<slra arguments for scalar Hankel problem>
p = []; for i = 1:N, p = [p; y(ind(i):ind(i) + T(i) - 1)]; end
[ph_m, info_m] = slra(p, s, r, opt);
```

The result is validated by computing the angle

$$\angle(\hat{R}, \bar{R}) := \min \left( \left| \cos^{-1} \frac{\hat{R}^\top \bar{R}}{\|\hat{R}\|_2 \|\bar{R}\|_2} \right|, \left| 180 - \cos^{-1} \frac{\hat{R}^\top \bar{R}}{\|\hat{R}\|_2 \|\bar{R}\|_2} \right| \right)$$

```
<define angle>≡
angle = @(a, b) min(abs(acosd(a(:)' * b(:) / norm(a) / norm(b))), ...
                    abs(180 - acosd(a(:)' * b(:) / norm(a) / norm(b))));
```

between the identified model parameter  $\hat{R}$  and the true model parameter  $\bar{R}$

```
<example autonomous SYSID>+≡
<define angle>, angle(info_m.Rh, R0)
```

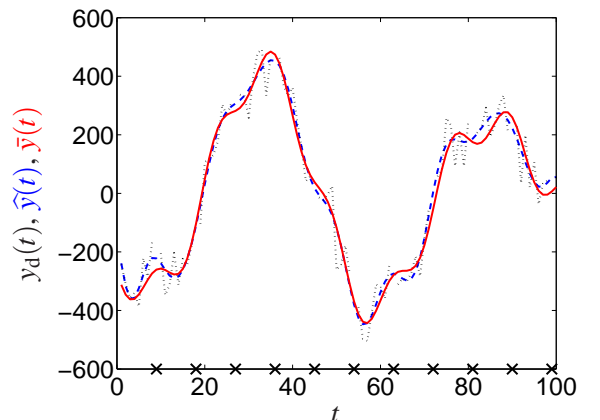
The result is  $4.7^\circ$ . Alternatively, the system is identified from the independent datasets with lengths  $T_1, \dots, T_N$ . The identified model parameters are also validated by their angles to the true model parameter.

```
<example autonomous SYSID>+≡
s = rmfield(s, 'n');
for i = 1:N
    p = y(ind(i):ind(i) + T(i) - 1);
    [ph, info] = slra(p, s, r, opt);
    angle(info.Rh, R0)
end
```

The obtained results are  $74^\circ$ ,  $18.5^\circ$ , and  $18.4^\circ$ , which shows the benefit of using all the data in a single approximation problem.

The following experiment shows identification with periodically missing data, in which case standard identification methods are not applicable.

```
<example autonomous SYSID>+≡
T = 100; Im = (ell + 1):(ell + 1):T;
p = y(1:100); p0 = y0(1:100); p(Im) = 0;
<s.w for missing values>
[ph_m, info_m] = slra(p, s, r, opt);
p(Im) = NaN;
figure(3), plot(ph_m, 'b-'), hold on,
plot(p, 'k:'), plot(p0(1:T), 'r-')
ax = axis;
plot(Im, ax(3) * ones(size(Im)), ...
     'Xk', 'markersize', 15)
axis(ax), print_fig('slra-f3')
```





## 5.5 Performance comparison with the old version of the package

In [32] and [36], the performance of the old version of the package is tested on benchmark system identification problems from the database DAISY [6]. The problems involve vector time-series and are fitted by multiple-inputs multiple-outputs linear time-invariant systems. The identification problem in this case is equivalent to block-rank low-rank approximation, which is also a special case of the mosaic Hankel low-rank approximation problem (SLRA). For a description of the test examples, we refer the reader to [32].

In Table 3, the execution time of the new and the old version of the package are compared for the same setting of the experiment as in [32]. In all examples, the approximate solutions computed by the two versions of the software are the same. The results show that the efficiency of the new version of the package is improved by an average of 40%. Note that the speedup is achieved by software improvements despite of the fact that the new version treats a more general problem.

#	Data set name	$T$	m	p	$\ell$	$t_{\text{new}}$ (sec)	$t_{\text{old}}$ (sec)	$\Delta t$ (%)
1	Data of a simulation of the western basin of Lake Erie	57	5	2	1	0.01	0.01	0
2	Data of ethane-ethylene distillation column	90	5	3	1	0.02	0.03	33
3	Heating system	801	1	1	2	0.01	0.02	50
4	Data from an industrial dryer (Cambridge Control Ltd)	867	3	3	1	0.16	0.25	36
5	Data of a laboratory setup acting like a hair dryer	1000	1	1	5	0.02	0.04	50
6	Data of the ball-and-beam setup in SISTA	1000	1	1	2	0.01	0.02	50
7	Wing flutter data	1024	1	1	5	0.02	0.04	50
8	Data from a flexible robot arm	1024	1	1	4	0.01	0.01	0
9	Data of a glass furnace (Philips)	1247	3	6	1	2.88	4.41	35
10	Heat flow density through a two layer wall	1680	2	1	2	0.09	0.22	59
11	Simulation data of a pH neutralization process	2001	2	1	6	0.03	0.06	50
12	Data of a CD-player arm	2048	2	2	1	0.09	0.23	61
13	Data from a test setup of an industrial winding process	2500	5	2	2	0.64	0.92	30
14	Liquid-saturated steam heat exchanger	4000	1	1	2	0.03	0.07	57
15	Data from an industrial evaporator	6305	3	3	1	1.36	2.22	39
16	Continuous stirred tank reactor	7500	1	2	1	0.24	0.75	68
17	Model of a steam generator at Abbott Power Plant	9600	4	4	1	13.10	15.77	17

Table 3: Performance test on examples from DAISY:  $T$ —number of data points,  $m$ —number of inputs,  $p$ —number of outputs,  $\ell$ —lag of the identified model,  $t_{\text{new}}$  and  $t_{\text{old}}$  execution times in seconds for the new and old versions of the package, respectively,  $\Delta t$ —percentage speedup, achieved by the new version of the package.

## 5.6 Approximate greatest common divisor of two polynomials

Consider the polynomials  $p^1$  and  $p^2$  of degrees  $d_1$  and  $d_2$ , respectively, and a positive integer  $\ell < \min(d_1, d_2)$ . The Sylvester matrix of  $p^1$  and  $p^2$  with parameter  $\ell$  is a

$$(d_1 + d_2 - 2\ell + 2) \times (d_1 + d_2 - \ell + 1)$$

mosaic Hankel matrix with upper-left and lower-right triangles of the block-elements fixed to zero:

$$S_\ell(p^1, p^2) := \begin{bmatrix} S_{d_2-\ell}(p^1) \\ S_{d_1-\ell}(p^2) \end{bmatrix}, \quad \text{where} \quad S_d(p) := \begin{bmatrix} p_0 & p_1 & \cdots & p_{d_p} \\ \cdot & \cdot & \cdot & \cdot \\ p_0 & p_1 & \cdots & p_{d_p} \end{bmatrix} \in \mathbb{R}^{(d+1) \times (d+d_p+1)}. \quad (\text{SYLV})$$

The degree of the greatest common divisor of  $p^1$  and  $p^2$  is equal to

$$\text{degree}(\gcd(p^1, p^2)) = d_1 + d_2 - \text{rank}(S_0(p^1, p^2)). \quad (*)$$

The considered approximate common divisor problem is defined as follows: given polynomials  $p^1$  and  $p^2$  and a lower bound  $\ell > 0$  on the degree of the common divisor, modify  $p^1$  and  $p^2$ , as little as possible, so that the modified polynomials  $\hat{p}^1$  and  $\hat{p}^2$  have a greatest common divisor of degree at least  $\ell$ , *i.e.*,

$$\begin{aligned} & \text{minimize} \quad \text{over } \hat{p}^1 \in \mathbb{R}^{d_1+1} \text{ and } \hat{p}^2 \in \mathbb{R}^{d_2+1} \quad \left\| \begin{bmatrix} p^1 \\ p^2 \end{bmatrix} - \begin{bmatrix} \hat{p}^1 \\ \hat{p}^2 \end{bmatrix} \right\|_2^2 \\ & \text{subject to} \quad \text{degree}(\gcd(\hat{p}^1, \hat{p}^2)) \geq \ell. \end{aligned} \quad (\text{ACD})$$

The approximate common divisor for the polynomials  $p^1$  and  $p^2$  is the exact greatest common divisor of  $\hat{p}^1$  and  $\hat{p}^2$ .

It can be shown [33] that the approximate common divisor problem (ACD) is equivalent to a Sylvester low-rank approximation problem

$$\begin{aligned} & \text{minimize} \quad \text{over } \hat{p}^1 \in \mathbb{R}^{d_1+1} \text{ and } \hat{p}^2 \in \mathbb{R}^{d_2+1} \quad \left\| \begin{bmatrix} p^1 \\ p^2 \end{bmatrix} - \begin{bmatrix} \hat{p}^1 \\ \hat{p}^2 \end{bmatrix} \right\|_2^2 \\ & \text{subject to} \quad \text{rank}(S_\ell(p^1, p^2)) \leq d_1 + d_2 - 2\ell + 1. \end{aligned}$$

```
<slra arguments for approximate GCD of 2 polynomials>≡
s.m = [d2 - ell + 1; d1 - ell + 1];
s.n = d1 + d2 - ell + 1;
z1 = zeros(d2 - ell, 1); z2 = zeros(d1 - ell, 1);
p = [z1; p1(:); z1; z2; p2(:); z2];
s.w = 1 ./ p; s.w(~isinf(s.w)) = 1;
r = d1 + d2 - 2 * ell + 1;
```

## Numerical examples

Our first example is Example 4.1 from [55]. The given polynomials are

$$\begin{aligned} p^1(z) &= (4 + 2z + z^2)(5 + 2z) + 0.05 + 0.03z + 0.04z^2 \\ p^2(z) &= (4 + 2z + z^2)(5 + z) + 0.04 + 0.02z + 0.01z^2 \end{aligned}$$

and an approximate common divisor of degree  $\ell = 2$  is sought.

```
<example GCD>≡
clear all; d1 = 3; d2 = 3; ell = 2;
p1 = conv([4 2 1], [5 2]) + [0.05 0.03 0.04 0];
p2 = conv([4 2 1], [5 1]) + [0.04 0.02 0.01 0];
<slra arguments for approximate GCD of 2 polynomials>
[ph, info] = slra(p, s, r);
```

The solution computed by the `slra` function

```
<example GCD>+≡
ph1 = ph(2:5), ph2 = ph(8:11), r_ph1 = roots(ph1), r_ph2 = roots(ph2)
```

ph1 =	ph2 =	r_ph1 =	r_ph2 =
20.0500	20.0392	-0.2510 + 0.4336i	-0.2510 + 0.4336i
18.0332	14.0179	-0.2510 - 0.4336i	-0.2510 - 0.4336i
9.0337	7.0176	-0.3973	-0.1974
2.0000	0.9933		

coincides (up to errors due to the numerical precision) with the one reported in [55].

The second example is Example 4.2, case 1, from [55] (originally given in [23]). The given polynomials are

$$\begin{aligned} p^1(\xi) &= (1 - \xi)(5 - \xi) = 5 - 6\xi + \xi^2 \\ p^2(\xi) &= (1.1 - \xi)(5.2 - \xi) = 5.72 - 6.3\xi + \xi^2 \end{aligned}$$

and an approximate common divisor of degree  $\ell = 1$  (a common root) is sought.

```
<example GCD>+≡
d1 = 2; d2 = 2; ell = 1;
p1 = conv([1 -1], [5 -1]);
p2 = conv([1.1 -1], [5.2 -1]);
<slra arguments for approximate GCD of 2 polynomials>
[ph, info] = slra(p, s, r);
```

Again, the solution computed by the `slra` function

```
<example GCD>+≡
ph1 = ph(2:4), ph2 = ph(7:9), r_ph1 = roots(ph1), r_ph2 = roots(ph2)

ph1 =          ph2 =          r_ph1 =          r_ph2 =
  4.9994          5.7206          1.0046          0.9047
 -6.0029         -6.2971          0.1961          0.1961
  0.9850          1.0150
```

coincides (up to numerical computation errors) with the one reported in the literature.

## 5.7 Approximate greatest common divisor of $N$ polynomials

Consider  $N$  polynomials  $p^1, \dots, p^N$  of degrees  $d_1, \dots, d_N$ , respectively. Similarly to the two polynomials case, the degree of the greatest common divisor of  $p^1, \dots, p^N$  can be expressed as

$$\text{degree}(\gcd(p^1, \dots, p^N)) = d_1 + \dots + d_N - \text{rank}(S_\ell(p^1, \dots, p^N))$$

where  $S_\ell(p^1, \dots, p^N)$  is the generalized (multipolynomial) Sylvester matrix [21] for  $p^1, \dots, p^N$  with parameter  $\ell$

$$S_\ell(p^1, \dots, p^N) := \begin{bmatrix} S_{d_1-\ell}(p^2) & \cdots & S_{d_1-\ell}(p^N) \\ S_{d_2-\ell}(p^1) & & \\ & \ddots & \\ & & S_{d_N-\ell}(p^1) \end{bmatrix}. \quad (\text{GSYLV})$$

$S_\ell(p^1, \dots, p^N)$  can be constructed as  $\Phi \mathcal{H}_{\mathbf{m}, \mathbf{n}}$  with specification of exact (zero) elements. For example, with  $N = 3$  and  $d_1 = d_2 = d_3 = 2$ ,

$$\Phi = \begin{bmatrix} I_4 & & \\ & 0_2 & I_2 \end{bmatrix} \quad \text{and} \quad \mathcal{H}_{[2 \ 6], 6} = \begin{bmatrix} S_1(p^2) & S_1(p^3) \\ S_6(p^1) \end{bmatrix}.$$

```
<slra arguments for approximate GCD of 3 polynomials>≡
s.m = [d1 - ell + 1; d2 - ell + 1 + d1 + d3 - ell + 1];
s.n = d1 + d2 - ell + 1 + d1 + d3 - ell + 1;
s.phi = eye(sum(s.m));
s.phi((s.m(1) + d2 - ell + 2):(s.m(1) + d2 - ell + 1 + d1), :) = [];
z1 = inf * ones(d1 - ell, 1); z2 = inf * ones(s.m(2) - 1, 1);
s.w = [z1; ones(d2 + 1, 1); z1; ones(d3 + 1, 1); z1; z2; ones(d1 + 1, 1); z2];
r = s.m(1) + d2 - ell + 1 + d3 - ell + 1 - ell;
p = zeros(size(s.w)); p(s.w == 1) = [p1 p2 p3]';
```

The approximate common divisor problem for  $N$  polynomials is defined as a direct generalization of (ACD):

$$\begin{aligned} & \text{minimize} \quad \text{over } \hat{p}^1 \in \mathbb{R}^{d_1+1}, \dots, \hat{p}^N \in \mathbb{R}^{d_N+1} \quad \left\| \begin{bmatrix} p^1 \\ \vdots \\ p^N \end{bmatrix} - \begin{bmatrix} \hat{p}^1 \\ \vdots \\ \hat{p}^N \end{bmatrix} \right\|_2^2 \\ & \text{subject to} \quad \text{degree}(\gcd(\hat{p}^1, \dots, \hat{p}^N)) \geq \ell \end{aligned}$$

and is equivalent to the low-rank approximation problem with rank constraint

$$\text{rank}(S_\ell(p^1, \dots, p^N)) \leq d_1 + \dots + d_N - N\ell + N - 1.$$

## Numerical examples

```
<example GCD>+=  
d3 = 2; p3 = conv([1.2 -1], [5.4 -1]);  
<slra arguments for approximate GCD of 3 polynomials>  
[ph, info] = slra(p, s, r);
```

```
<example GCD>+=  
ph123 = ph(s.w == 1);  
r_ph1 = roots(ph123(1:d1 + 1))  
r_ph2 = roots(ph123(d1 + 2:d1 + d2 + 2))  
r_ph3 = roots(ph123(d1 + d2 + 3:end))
```

r_ph1 =	r_ph2 =	r_ph3 =
1.0090	0.9090	0.8251
0.1924	0.1924	0.1924

## 6 Conclusions

The developed software package is a generic tool for data modeling and has numerous applications in system theory and identification, signal processing, machine learning, chemometrics, and computer algebra. Its functionality generalized the one of the software of [35] by allowing specification of element-wise weights, arbitrary fixed and missing elements, linear constraints on the parameter matrix, and by generalizing the structure of the approximating matrix to the class of mosaic Hankel-like matrices. Planned extensions of the package are

- multiple rank constraints,
- mosaic Hankel matrices with repeated Hankel blocks (*i.e.*, structure parameters common to two or more blocks),
- linear equality constraints on the structure parameter vector,
- nonlinear structure of the kernel parameter  $R$ .

On the practical side, we experiment with different optimization strategies (methods for optimization on a Grassmann manifold) and compare the performance of the package with state-of-the-art problem dependent methods for the applications in Table 2.

## Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement number 258581 "Structured low-rank approximation: Theory, algorithms, and applications".

## References

- [1] GSL - GNU scientific library. [www.gnu.org/software/gsl/](http://www.gnu.org/software/gsl/).
- [2] T. Abatzoglou, J. Mendel, and G. Harada. The constrained total least squares technique and its application to harmonic superresolution. *IEEE Trans. Signal Proc.*, 39:1070–1087, 1991.
- [3] A. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, 2005.
- [4] J.-F. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. [www-stat.stanford.edu/~candes/papers/SVT.pdf](http://www-stat.stanford.edu/~candes/papers/SVT.pdf), 2009.
- [5] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2009.
- [6] B. De Moor. DaISy: Database for the identification of systems. [www.esat.kuleuven.be/sista/daisy/](http://www.esat.kuleuven.be/sista/daisy/).

- [7] B. De Moor. Structured total least squares and  $L_2$  approximation problems. *Linear Algebra Appl.*, 188–189:163–207, 1993.
- [8] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial Mathematics, 1987.
- [9] S. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38:189–230, 2004.
- [10] G. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15:318–344, 1973.
- [11] G. Golub, P. Milanfar, and J. Varah. A stable numerical method for inverting shape from moments. *SIAM J. Sci. Comput.*, 21:1222–1243, 1999.
- [12] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Institute of Physics, Inverse Problems*, 19:1–26, 2003.
- [13] G. Golub and C. Van Loan. An analysis of the total least squares problem. *SIAM J. Numer. Anal.*, 17:883–893, 1980.
- [14] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.
- [15] R. Gorsuch. *Factor Analysis*. Hillsdale, 1983.
- [16] R. Guidorzi. Invariants and canonical forms for systems structural and parametric identification. *Automatica*, 17(1):117–133, 1981.
- [17] P. Guillaume and R. Pintelon. A Gauss–Newton-like optimization algorithm for “weighted” nonlinear least-squares problems. *IEEE Trans. Signal Process.*, 44(9):2222–2228, 1996.
- [18] G. Heinig. Generalized inverses of hankel and toeplitz mosaic matrices. *Linear Algebra Appl.*, 216(0):43–59, February 1995.
- [19] Markovsky I, J. C. Willems, S. Van Huffel, and B. De Moor. *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*. SIAM, March 2006.
- [20] T. Kailath and A. Sayed. *Fast reliable algorithms for matrices with structure*. SIAM, Philadelphia, 1999.
- [21] E. Kaltofen, Z. Yang, and L. Zhi. Approximate greatest common divisors of several polynomials with linearly constrained coefficients and singular polynomials. In *Proc. Int. Symp. Symbolic Algebraic Computation*, pages 169–176, 2006.
- [22] N. Karmarkar and Y. Lakshman. On approximate GCDs of univariate polynomials. In S. Watt and H. Stetter, editors, *J. Symbolic Comput.*, volume 26, pages 653–666, 1998.
- [23] N. Karmarkar and Y. Lakshman. On approximate GCDs of univariate polynomials. In S. Watt and H. Stetter, editors, *J. Symbolic Comput.*, volume 26, pages 653–666, 1998. Special issue on Symbolic Numeric Algebra for Polynomials.
- [24] R. Kumaresan and D. Tufts. Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise. *IEEE Trans. Acoust., Speech, Signal Proc.*, 30(6):833–840, 1982.
- [25] P. Lemmerling, N. Mastronardi, and S. Van Huffel. Fast algorithm for solving the Hankel/Toeplitz structured total least squares problem. *Numerical Algorithms*, 23:371–392, 2000.
- [26] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, NJ, 1999.
- [27] J. Manton, R. Mahony, and Y. Hua. The geometry of weighted low-rank approximations. *IEEE Trans. Signal Process.*, 51(2):500–514, 2003.
- [28] I. Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008.
- [29] I. Markovsky. Fast measurements of slow processes. Technical Report 22082, ECS, Univ. of Southampton, <http://eprints.ecs.soton.ac.uk/22082/>, 2011.
- [30] I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.
- [31] I. Markovsky and K. Usevich. Structured low-rank approximation with fixed and missing values. Technical report, ECS, Univ. of Southampton, <http://eprints.ecs.soton.ac.uk/>, 2012.
- [32] I. Markovsky and S. Van Huffel. High-performance numerical algorithms and software for structured total least squares. *J. Comput. Appl. Math.*, 180(2):311–331, 2005.
- [33] I. Markovsky and S. Van Huffel. An algorithm for approximate common divisor computation. In *Proc. 17th Symp. on Math. Theory of Networks and Systems*, pages 274–279, Kyoto, Japan, 2006.
- [34] I. Markovsky and S. Van Huffel. Overview of total least squares methods. *Signal Processing*, 87:2283–2302, 2007.

- [35] I. Markovsky, S. Van Huffel, and R. Pintelon. Block-Toeplitz/Hankel structured total least squares. *SIAM J. Matrix Anal. Appl.*, 26(4):1083–1099, 2005.
- [36] I. Markovsky, J. C. Willems, and B. De Moor. Comparison of identification algorithms on the database for system identification DAISY. In *Proc. 17th Symp. on Math. Theory of Networks and Systems*, pages 2858–2869, Kyoto, Japan, 2006.
- [37] I. Markovsky, J. C. Willems, S. Van Huffel, and B. De Moor. *Exact and approximate modeling of linear systems: A Behavioral Approach*. SIAM, March 2006.
- [38] I. Markovsky, J. C. Willems, S. Van Huffel, B. De Moor, and R. Pintelon. Application of structured total least squares for system identification and model reduction. *IEEE Trans. Automat. Control*, 50(10):1490–1500, 2005.
- [39] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.
- [40] N. Mastronardi, P. Lemmerling, and S. Van Huffel. Fast structured total least squares algorithm for solving the basic deconvolution problem. *SIAM J. Matrix Anal. Appl.*, 22:533–553, 2000.
- [41] P. Milanfar, G. Verghese, W. Karl, and A. Willsky. Reconstructing polygons from moments with connections to array processing. *IEEE Trans. Signal Proc.*, 43:432–443, 1995.
- [42] R. Pintelon and J. Schoukens. Identification of continuous-time systems with missing data. *IEEE Trans. Instr. Meas.*, 48(3):736–740, 1999.
- [43] O. Rojo. A new algebra of Toeplitz-plus-Hankel matrices and applications. *Computers & Mathematics with Appl.*, 55(12):2856–2869, 2008.
- [44] J. Rosen, H. Park, and J. Glick. Total least norm formulation and solution of structured problems. *SIAM J. Matrix Anal. Appl.*, 17:110–126, 1996.
- [45] T. Segaran. *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O’Reilly Media, 2007.
- [46] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.
- [47] P. Stoica and R. Moses. *Spectral Analysis of Signals*. Prentice Hall, 2005.
- [48] K. Usevich and I. Markovsky. Structured low-rank approximation as a rational function minimization. In *Proc. of the 16th IFAC Symposium on System Identification*, Brussels, 2012.
- [49] D. van Heesch. Doxygen. [www.doxygen.org](http://www.doxygen.org).
- [50] S. Van Huffel, V. Sima, A. Varga, S. Hammarling, and F. Delebecque. High-performance numerical software for control. *IEEE Control Systems Magazine*, 24:60–76, 2004.
- [51] S. Van Huffel and J. Vandewalle. *The total least squares problem: Computational aspects and analysis*. SIAM, Philadelphia, 1991.
- [52] P. Wentzell, D. Andrews, D. Hamilton, K. Faber, and B. Kowalski. Maximum likelihood principal component analysis. *J. Chemometrics*, 11:339–366, 1997.
- [53] J. C. Willems. From time series to linear system—Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling. *Automatica*, 22, 23:561–580, 675–694, 87–115, 1986, 1987.
- [54] J. C. Willems. The behavioral approach to open and interconnected systems: Modeling by tearing, zooming, and linking. *Control Systems Magazine*, 27:46–99, 2007.
- [55] L. Zhi and Z. Yang. Computing approximate GCD of univariate polynomials by structure total least norm. In *MM Research Preprints*, number 24, pages 375–387. Academia Sinica, 2004.

## A Implementation details

The structured low-rank approximation solver is written in C++ language and the implementation uses object-oriented style. The LAPACK and BLAS libraries are used for the matrix computations and the GNU scientific library (GSL) [1] is used for vector-matrix multiplications and for nonlinear optimization. The key computational step—Cholesky factorization of the  $\Gamma$  matrix—is done by the LAPACK’s function `DPBTRF`, which exploits the banded structure of the matrix. In case of blockwise weights (see Section 3), the package can optionally use the `MB02GD` function from the SLICOT library [50], which exploits both the banded and Toeplitz structure of  $\Gamma$ . (In our test examples `MB02GD` gives no advantage in terms of computation time.)

The package is divided into the following sub directories:



- `cpp`: C++ core classes and functions.<sup>2</sup>
- `mex`: source files for Matlab/Octave interface.
- `Rslra`: source files for R interface.
- `test_m`: demo files for Matlab/Octave interface.
- `test_r`: demo files for R interface.
- `test_c`: demo files for C++ interface.
- `doc`: documentation and examples.
- `ident`: wrapper function for system identification.

## A.1 Main function

The solver is called in C via the function `slra`, defined as follow:

```
<slra function definition in C>≡
int slra( const gsl_vector *p_in, Structure *s, int d, OptimizationOptions *opt,
          gsl_matrix *Rini, gsl_matrix *Phi, gsl_matrix *Psi,
          gsl_vector *p_out, gsl_matrix *r_out, gsl_matrix *v_out );
```

- `p_in`: input vector  $p$ .
- `s`: object of type `Structure`, see Section A.2.
- `d`: rank reduction.
- `opt`: optimization options and output information.
- `Rini`: transposed initial approximation.
- `Phi`:  $\Phi^\top$  matrix (see  $(\mathcal{S})$ ).
- `Psi`:  $\Psi^\top$  matrix (see  $(\Theta \mapsto R)$ ).
- `p_out`: approximation  $\hat{p}$ .
- `R_out`: low-rank certificate  $\hat{R}^\top$ .
- `v_out`: error covariance  $(J^\top J)^{-1}$  of  $\text{vec}(X)$ .

*Note 8* (C row-major convention). All matrices are transposed, due to the row-major order convention of C/GSL.

## A.2 Structure specification and object-oriented paradigm

The `Structure` class represents the structure specification  $\mathcal{S}$  and the weights vector  $w$ .

```
<definition of class Structure>≡
class Structure {
public:
    virtual ~Structure() {}
    virtual int getNp() const = 0;
    virtual int getM() const = 0;
    virtual int getN() const = 0;
    virtual void fillMatrixFromP( gsl_matrix* c, const gsl_vector* p ) = 0;
    virtual void correctP( gsl_vector* p, gsl_matrix *R, gsl_vector *yr,
                          bool scaled = true ) = 0;
    virtual Cholesky *createCholesky( int D, double reg_gamma ) const = 0;
    virtual DGamma *createDGamma( int D ) const = 0;
};
```

The function `slra` can deal with an arbitrary `Structure` object using the following methods:

- `getNp()`, `getM()` and `getN()` return  $n_p$ ,  $m$  and  $n$ , respectively.
- `fillMatrixFromP()` constructs the matrix  $\mathcal{S}(p)^\top$  from the vector of structure parameters  $p$ .
- `correctP()` constructs the correction vector  $\Delta p$  from  $R$  and precomputed  $y_r := \Gamma^{-1}(R)v(R)$ .
- `createCholesky()` and `createDerivative()` create objects for Cholesky factorization and gradient/Jacobian computation respectively.

The currently implemented structures are:

---

<sup>2</sup>Type declarations and function prototypes are in `slra_XXX.h` files; the implementation is in `slra_XXX.cpp` files.

- `LayeredHStructure`: class for layered Hankel structure with blockwise weights,
- `WLayeredHStructure`: class for layered Hankel structure with elementwise weights, and
- `MosaicHStructure` and `WMosaicHStructure`: classes for mosaic Hankel structure, that are implemented based on layered Hankel structure.

The object-oriented paradigm facilitates the memory management and software design. In particular, it allows the user to add a new problem specification (for example, a new type of structure) by implementing a new `Structure` subclass. Further details can be found in the manual for the C++ interface, which can be generated by running `doxygen` [49] in the `cpp` directory.

### A.3 Installation instructions

The package is distributed in the form of source code. The interfaces for Matlab/Octave and R can be compiled by calling `make` in the root directory of the package. The GSL, BLAS, and LAPACK libraries have to be installed in advance. The SLICOT library can be used optionally.

## B Total least squares vs low-rank approximation

The total least squares problem

$$\text{minimize over } X, \hat{A}, \text{ and } \hat{B} \quad \left\| \begin{bmatrix} A & B \end{bmatrix} - \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} \right\|_F \quad \text{subject to} \quad \hat{A}X = \hat{B} \quad (\text{TLS})$$

was introduced by Golub and Van Loan [10, 13] as a solution technique for an overdetermined system of linear equations  $AX \approx B$ . With  $X \in \mathbb{R}^{r \times \bullet}$ , (TLS) is generically equivalent to approximation of the matrix  $D := \begin{bmatrix} A & B \end{bmatrix}$  by a rank  $r$  matrix  $\hat{D} := \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix}$  in the Frobenius norm. While a solution to the low-rank approximation problem

$$\text{minimize over } \hat{D} \in \mathbb{R}^{m \times n} \quad \|D - \hat{D}\|_F^2 \quad \text{subject to} \quad \text{rank}(\hat{D}) \leq r. \quad (\text{LRA})$$

always exists, the total least-squares problem may fail to have a solution. The nongeneric case of lack of total least squares solution occurs when the optimal approximating matrix  $\hat{D}$  of (LRA) can not be represented in the form  $\hat{A}X \approx \hat{B}$ .

The system of equations  $\hat{A}X \approx \hat{B}$  implies a functional relation among the variables  $\hat{A}$  to  $\hat{B}$  ( $\hat{B}$  is a function of  $\hat{A}$ ). In system theory and signal processing, such a relation defines what is called an input-output representation of a model. The rank constraint in (LRA) does not impose an a priori fixed functional relation or input-output partition on the variables of the matrix  $\hat{D}$ . (Of course, the rank deficiency of  $\hat{D}$  implies that such relations exist and they can be inferred from  $\hat{D}$  after the approximation problem is solved.) This representation-free approach to data modeling is advocated by J. C. Willems and is known in the systems and control literature as the behavioral paradigm [53].

The generalization of the total least squares problem to systems of linear equations with structured matrices  $A$  and  $B$  is called structured total least squares. As in the unstructured case, the structured total least squares is generically equivalent to structured low-rank approximation. Many algorithms for solving structured total least squares problems [2, 7, 44, 25, 40] are proposed in the literature. For more information about the (structured) total least squares approach and its applications, we refer the reader to the overview paper [34], proceedings and special issues and the books [51, 19].

Motivated by the advantages of the behavioral paradigm over the classical input-output setting, see [54], we adopt in the updated version of the software the low-rank approximation problem as the generic problem formulation.

## C Optional input arguments for the optimization method

The following optional input arguments of the `slra` function are related to the optimization solver, used for the solution of the parameter optimization problem. The options can be specified either as fields of the input argument `opt` or, when calling via the wrapper function, as pairs ('name', 'value') of input arguments.

- `solver` specifies the solver. The options are:
  - `c` — efficient C++ solver (default), and
  - `m` — general but inefficient solver, implemented in Matlab.
- `method` specifies the optimization method to be used with the C++ solver. The available options are all local optimization methods in the GSL library (see the GSL manual [1] for more details):
  - `l` — Levenberg–Marquardt methods for nonlinear least squares (default),
    - \* `ll` — method `lmder` (default),
    - \* `ls` — method `lmsder`,
  - `q` — Quasi-Newton and conjugate gradient methods for nonlinear minimization with derivatives,
    - \* `qb` — method `bfgs` (default),
    - \* `q2` — method `bfgs2`,
    - \* `qp` — method `conjugate_pr`,
    - \* `qf` — method `conjugate_fr`,
  - `n` — methods for nonlinear minimization without derivatives,
    - \* `nn` — method `nmsimplex` (default),
    - \* `n2` — method `nmsimplex2`, and
    - \* `nr` — method `nmsimplex2rand`.
- `disp` specifies the level of displayed information. The options are:
  - `iter` — print progress information per iteration,
  - `notify` — in case of lack of convergence only (default), or
  - `off` — no display.
- The following arguments control the termination of the optimization:
  - `maxiter` — maximum number of iterations,
  - `tol` — tolerance for the change of the cost function value,
  - `epsrel` and `epsabs` — relative and absolute tolerance for the element-wise change of the optimization variables,
  - `epsgrad` — tolerance for the norm of the gradient.