Authors' response to the referees' reports on "Structured low-rank approximation with missing data"

I. Markovsky and K. Usevich

We thank the referees for their relevant and useful comments. In this document, we quote in **bold face** statements from the referee reports. Our replies follow in ordinary print. In blue, we quote passages from the revised manuscript.

Answer to referee #1

1.1: Since this work builds on previous work on structured total least squares methods in [13], it should be clearly stated what the connection is. How is the (SLRA) problem different from [13] and why is it challenging? Is the only difference that there are missing values in (SLRA)?

The structured total least squares problem

$$\begin{array}{ll} \text{minimize} & \text{over } X \in \mathbb{R}^{r\times(m-r)}, \widehat{A} \in \mathbb{R}^{n\times r}, \text{ and } \widehat{B} \in \mathbb{R}^{n\times(m-r)} \quad \left\| \begin{bmatrix} A & B \end{bmatrix} - \begin{bmatrix} \widehat{A} & \widehat{B} \end{bmatrix} \right\|_{\mathrm{F}} \\ \text{subject to} & \widehat{A}X = \widehat{B} \quad \text{and} \quad \begin{bmatrix} \widehat{A} & \widehat{B} \end{bmatrix} \text{ is structured} \end{array} \tag{TLS}$$

is a restriction of the structured low-rank approximation problem

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{D} \in \mathbb{R}^{m\times n} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \\ \text{subject to} & \operatorname{rank}(\widehat{D}) \leq r \quad \text{and} \quad \widehat{D} \text{ is structured} \end{array} \tag{LRA}$$

to the subclass of rank r matrices

$$\widehat{D} = \begin{bmatrix} \widehat{A} & \widehat{B} \end{bmatrix}^\top, \qquad \text{such that} \quad \widehat{A}X = \widehat{B}.$$

The similarity between the solution technique presented in the paper under review (called "the paper") and the one of

> [MVP05] 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

is in the use of the variable projections method

> 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

for the elimination of the optimization variables $\widehat{A}$ and $\widehat{B}$ in (TLS) and $\widehat{D}$ in (LRA).

Our experience with the variable projection approach helped us in deriving the new results, however, the following points are important differences between [MVP05] and the paper.

- After the elimination step, the remaining nonlinear optimization problem is over a the set of full row rank matrices $R \in \mathbb{R}^{(m-r)\times m}$ instead of the set of unconstrained matrices $X \in \mathbb{R}^{r\times(m-r)}$. As a consequence, the optimization problem in the paper is on a Stiefel manifold, while in [MVP05] it is on an Euclidean space.
- In [MVP05], missing values are not considered, while the focus of the paper reviewed is on the missing values.
- The main result in [MVP05] is a fast method for Hankel structured problems, while in the paper, the structure is general affine and efficient computation in the case of Hankel structure is not considered.

- In [MVP05], the approximation is in the two norm, while in the paper general weighted 2-norm approximation is allowed.

On one hand, the link between total least squares and low-rank approximation is well documented in the literature, see, *e.g.*, the survey papers

> I. Markovsky and S. Van Huffel. Overview of total least squares methods. *Signal Proc.*, 87:2283–2302, 2007
>
> I. Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008

and the book

> I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer, 2012

On another hand, the generalization of the problem and the differences above make our new results rather remote from the results in [MVP05]. For this reason we decided to de-emphasize the role of [MVP05] and include only the following sentence about the link of new results to the results on the Hankel structural total least squares problem:

... we aim to unify as many data modeling applications as possible and derive a single algorithm that solves them. Of course, this goal can be achieved by brute force optimization. The challenge is to discover and use effectively the structure of the problem. In the general problem considered, this structure is a separation of variables with analytic solution over one set of variables. This approach is related to the variable projections method [GP03] used in [MVP05].

1.2: The readability of the abstract could be improved. There are too many technical terms which are not defined yet, such as "structural total least squares', "weighted structured low-rank approximation', and "elimination of correction matrix'.

In order to avoids the use of the technical terms, the abstract is rewritten as follow:

We consider low-rank approximation of affinely structured matrices with missing elements. The method proposed is based on reformulation of the problem as inner and outer optimization. The inner minimization is a singular linear least-norm problem and admits an analytic solution. The outer problem is a nonlinear least squares problem and is solved by local optimization methods: minimization subject to quadratic equality constraints and unconstrained minimization with regularized cost function. The method is generalized to weighted low-rank approximation with missing values and is illustrated on approximate low-rank matrix completion, system identification, and data-driven simulation problems. An extended version of the paper is a literate program, implementing the method and reproducing the presented results.

1.3: In the second paragraph of Section 1, do not mention 'structural total least squares method' if it is not explained in this paragraph. There is no reason to state the last two sentences of the second paragraph, unless the authors can support the claim that 'the subject is closely related to the structural total least squares method'.

See the answer to question 1.1.

1.4: In the third paragraph of Section 1, 'the unstructured low-rank approximation problem with missing data' is commonly known as 'matrix completion problem' or 'low-rank matrix completion problem'. The readers who are more familiar with the term 'matrix completion' could benefit from making this explicit. Also, add references to more recent work on matrix completion such as 'exact matrix completion via convex optimization' by Candes and Recht.

Indeed, in the case of unstructured matrix, the problem treated in the paper is approximate low-rank matrix completion. We have added the following explanation of the link to the matrix completion problem in subsection "Applications".

Linear static modeling with missing data: An approximate low-rank matrix completion problem

Consider a set of vectors $\mathscr{D} = \{d_1, \ldots, d_N\}$ in $\mathbb{R}^q$. A linear model $\mathscr{B}$ for the data $\mathscr{D}$ is a subspace of the data space $\mathbb{R}^q$, and the dimension of $\mathscr{B}$ is a measure of the model's complexity. Linear static modeling is the problem of finding a low-complexity model (low-dimensional subspaces) that fits the data as close as possible. Existence of an exact linear model $\mathscr{B}$ for the data $\mathscr{D}$, *i.e.*, $\mathscr{D} \subset \mathscr{B}$, with complexity at most r is equivalent to the data matrix $D = \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}$ having rank at most r. Therefore, measuring the fit between the data point d_i and the model $\mathscr{B}$ by the orthogonal distance

$$\operatorname{dist}(d_i, \mathscr{B}) = \min_{\widehat{d}_i \in \mathscr{B}} \|d_i - \widehat{d}_i\|_2^2,$$

the approximate fitting problem becomes a rank r matrix approximation problem (SLRA) with unstructured data matrix $\mathscr{S}(p) = D$.

Suppose now that some elements d_{ij}, $(i,j) \in \mathscr{I}_{\text{missing}}$ of the data matrix are missing. Equivalently, only the elements d_{ij}, $(i,j) \in \mathscr{I}_{\text{given}}$ of D are specified. The exact linear static modeling problem becomes a low-rank matrix completion problem [CR09],

$$\text{find} \quad \widehat{D} \quad \text{such that} \quad \operatorname{rank}(\widehat{D}) \leq r \text{ and } \widehat{D}_{\mathscr{I}_{\text{given}}} = D_{\mathscr{I}_{\text{given}}}.$$

Here $D_{\mathscr{I}}$ denotes the vector of elements of D with indeces in $\mathscr{I}$. In the context of approximate data fitting by linear static model and missing data values, the relevant problem is approximate low-rank matrix completion [CP09]:

$$\text{minimize} \quad \text{over } \widehat{D} \quad \|\widehat{D}_{\mathscr{I}_{\text{given}}} - D_{\mathscr{I}_{\text{given}}}\|_2^2 \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r. \tag{AMC}$$

The approximate low-rank matrix completion problem (AMC) is used for building recommender systems. In recommender system applications, there is a set of users and a set of products. Some users rate some products. The goal is to predict the user ratings on products that they have not rated. The underlying assumption that makes the solution of this problem possible is that the full user-ratings matrix is low-rank. The low-rank property is observed empirically and can be explained intuitively as existence of a small number of groups of users with the same "taste" (*i.e.*, users that like or dislike the same products). In practice, the low-rank assumption is satisfied only approximately, which makes the approximation aspect of the problem essential.

The main issue in building real-life recommender systems is the high dimensionality and sparsity of the data matrix. Additional important issues in building practical recommender systems is the fact that the given and missing ratings are discrete and that apart from the users' ratings, there is demographic information about the users. Taking into account this prior information may improve significantly the accuracy of the missing values estimation. These issues, however, are outside the scope of the present paper.

1.5: The readers could benefit from adding a section for related work. From applications point of view, it would be useful to have explain in more detail the problems ... with appropriate back ground and references to related literature. Making this connection (why are these problems SLRA problem?) explicit will improve the relevance of this paper greatly.

The new subsection "Applications" of the introduction addresses this comment.

1.6: (a) matrix completion

See, the answer to question 1.4.

1.7: (b) system identification

The following description of the link between system identification and low-rank approximation is added.

System identification with missing data: An approximate Hankel structured low-rank matrix completion problem

A discrete-time linear time-invariant dynamical model is a set of time series

$$\mathscr{B}(R) := \left\{ w \mid R_0 w(t) + R_1 w(t+1) + \cdots + R_\ell w(t+\ell) = 0, \text{ for all } t \right\} \tag{$\mathscr{B}$}$$

that satisfy a constant coefficients difference equation. The matrices $R_0, R_1, \ldots, R_\ell \in \mathbb{R}^{\mathtt{p} \times q}$ are parameters specifying the model. Note that the linear static model is a special case of a linear time-invariant dynamical model when the lag ℓ of the difference equation representing the system is equal to zero.

A finite time series

$$w = \big(w(1), \ldots, w(T)\big), \qquad \text{where} \quad w(t) \in \mathbb{R}^q,$$

is an exact trajectory of the system defined in ($\mathscr{B}$) if the following matrix equation is satisfied

$$\underbrace{\begin{bmatrix} R_0 & R_1 & \cdots & R_\ell \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} w(1) & w(2) & w(3) & \cdots & w(T-\ell) \\ w(2) & w(3) & \iddots & & w(T-\ell+1) \\ w(3) & \iddots & & & \vdots \\ \vdots & & & & \\ w(\ell+1) & w(\ell+2) & \cdots & & w(T) \end{bmatrix}}_{\mathscr{H}_{\ell+1}(w)} = 0.$$

Without loss of generality, we assume that the parameter matrix R has full row rank $\mathtt{p}$, which implies that

$$\operatorname{rank}\big(\mathscr{H}_{\ell+1}(w)\big) \leq q(\ell+1) - \mathtt{p}.$$

We showed above that the data w is an exact trajectory of a system $\mathscr{B}(R)$, if the Hankel matrix $\mathscr{H}_{\ell+1}(w)$ is rank deficient. Therefore, as in the static case, approximate modeling by a linear time-invariant system is a low-rank approximation problem

$$\text{minimize} \quad \text{over } \widehat{w} \quad \|w - \widehat{w}\|_2^2 \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{H}_{\ell+1}(\widehat{w})\big) \leq q(\ell+1) - \mathtt{p}. \tag{SYSID}$$

Note, however, that the linear time-invariant model class imposes a Hankel structure constraint on the approximation matrix.

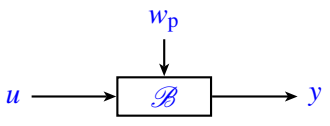
Identification from a trajectory with missing elements is therefore a Hankel structured low-rank matrix completion problem. A special system identification problems for the class of auto-regressive exogenous (ARX) systems with missing data is considered in [Isa93], a method based on frequency domain techniques is proposed in [PS00], and a method combining subspace identification and nuclear norm minimization is presented in [LHV12]. These papers do not link the system identification problem to the Hankel low-rank approximation problem (SYSID), so that their approaches are different from ours. The method developed in this paper, when specialized to block-Hankel structure can be used for general multivariable system identification in the time domain.

1.8: (c) data-driven simulation

The following description of the link between data-driven simulation and low-rank approximation is added.

Data-driven simulation and control

The trajectory w of a dynamical system can be partitioned into inputs u, *i.e.*, free variables, and outputs y, *i.e.*, variables that are determined by the inputs, initial conditions, and the model. Let $w = (u, y)$ be such a partition. The output $y = \big(y(1), \ldots, y(T)\big)$ of $\mathscr{B}$ is uniquely determined by the input $u = \big(u(1), \ldots, u(T)\big)$ and the initial conditions



$$w_{\mathrm{p}} = \big(w(-\ell+1), w(-\ell+1), \ldots, w(0)\big).$$

This gives a "signal processor" interpretation of a dynamical system.

The simulation problem aims to find the output y_{f} of a system $\mathscr{B}$, corresponding to a given input u_{f} and initial conditions w_{p}, *i.e.*,

$$\text{find} \quad y_{\mathrm{f}} \quad \text{such that} \quad w_{\mathrm{p}} \wedge (u_{\mathrm{f}}, y_{\mathrm{f}}) \in \mathscr{B}.$$

($w_{\mathrm{p}} \wedge w_{\mathrm{f}}$ denotes the concatenation of w_{p} and w_{f}.) This is a classic problem in system theory and numerical linear algebra, for which many solutions exist, *e.g.*, for systems with no inputs, the problem is related to the computation of the matrix exponential [MV78]. The classical simulation methods require a representation (state space, transfer function, convolution kernel, *etc*.) of the model. Such a representation is often obtained from data by system identification. The question occurs of solving the simulation problem directly from the data without identifying a model representation as a byproduct and using it in a model based solution. We call the direct problem data-driven simulation [MR08].

The data-driven simulation problem is a mosaic-Hankel structured low-rank approximation problem with missing data. To see this, let w' denotes the data that implicitly specifies the model and $w'' = w''_{\mathrm{p}} \wedge (u''_{\mathrm{f}}, y''_{\mathrm{f}})$ be the to-be-simulated trajectory. We express the condition that w' and w'' are trajectories of a linear time-invariant system with lag ℓ in matrix language as

$$\operatorname{rank}\left(\begin{bmatrix} \mathscr{H}_{\ell+1}(w') & \mathscr{H}_{\ell+1}(w'') \end{bmatrix}\right) \leq q(\ell+1) - \mathtt{p}.$$

This is a model-free description of the simulation problem—the existence of a model is implicit in the rank constraint. The missing data are the to-be computed response y_{p}. When the data w' is not an exact trajectory of the model, the matrix $\mathscr{H}_{\ell+1}(w')$ is generically full rank, so that an approximation is needed. The resulting data-driven simulation problem is a Hankel structured low-rank approximation with missing data:

$$\begin{array}{lll} \text{minimize} & \text{over } \widehat{w}' \text{ and } \widehat{w}'' & \|w' - \widehat{w}'\|_2^2 \\ \text{subject to} & \operatorname{rank}\left(\begin{bmatrix} \mathscr{H}_{\ell+1}(\widehat{w}') & \mathscr{H}_{\ell+1}(\widehat{w}'') \end{bmatrix}\right) \leq q\ell + \mathtt{m}, & \\ & \widehat{w}''_{\mathrm{p}} = w''_{\mathrm{p}}, \quad \text{and} \quad \widehat{u}''_{\mathrm{f}} = u''_{\mathrm{f}}. & \end{array} \tag{DDSIM}$$

1.9: Also, another paragraph or section on related work in (SLRA) problem would be helpful.Is this the first paper addressing this problem? If so, what is the main challenge? How is it different from similar problems that has been studied? Is it the 'missing data' part that makes this problem new? Or is it the low-rank approximation that has not been addressed in previous literature in structured matrix approximation problems?

- *Novelty:* To the best of our knowledge this is the first paper dealing with missing data in the context of structured low-rank approximation. Because of the generality of the problem, special cases appear frequently in applications, however, they are often viewed as problems in the particular application context. The originality and main contribution of the reported work is in the unification of many problems in systems and control, signal processing, machine learning, and computer algebra.

- *Challenges:* As in the unstructured matrix completion problem, the main challenge is the nonconvexity of the problem. We use local optimization heuristics. The challenge with this approach is to discover and use the problem structure. In the problem at hand we exploit the bilinear structure of the constraint $R\mathscr{S}(\widehat{p}) = 0$, which leads to the analytic solution of the minimization problem with respect to $\widehat{p}$ and therefore reduces the problem to minimization over R only. In the case of missing data, this step is a singular least-norm problem. Another challenge (left for future work) is fast cost function and derivatives evaluation in the special case of mosaic-Hankel structure with missing values. Fast algorithms for mosaic-Hankel problems without missing values are presented in

  > K. Usevich and I. Markovsky. Variable projection for affinely structured low-rank approximation in weighted 2-norm, 2012. Available from `http://arxiv.org/abs/1211.3938`

  The generalization of these methods to the missing data case is not straightforward and requires extra work.

The following explanation is added in the revised version of the paper:

Although structured low-rank approximation and approximate low-rank matrix completion (missing data estimation in low-rank matrices) are independently active research topics, the combined problem of missing data estimation in affine structured low-rank matrices has not been considered before. Both the structured low-rank approximation and approximate matrix completion problems are nonconvex optimization problems that in general admit no analytic solution. Therefore, in both domains local optimization and convex relaxation heuristics are used as solution techniques. In this paper, we use the local optimization approach.

Structured low-rank approximation has been studied in the literature from different viewpoints: numerical algorithm for computing locally optimal or suboptimal solutions, statistical properties of the resulting estimators, and applications. From a numerical point of view, the main challenge is to achieve fast and robust computational methods that can deal effectively with large data sets. From a statistical point of view, the main challenge is to establish conditions for consistency and efficiency of the methods. Our objective in this paper is different: we aim to unify as many applications as possible and derive a single algorithm that solves them. Of course, this goal can be achieved by brute force optimization. The challenge is to discover and use effectively the structure of the problem. In the general problem considered, this structure is a separation of variables with analytic solution over one set of variables. Our approach of exploiting this structure is related to the variable projections method [GP03], used in [MVP05].

1.10: Numerical experiments in Section 5. should be improved. In section 5.1, a 3×10 matrix is used for this problem. A typical size of numerical experiments in other matrix completion literature would be 1000×1000 matrices. The given example is too small to make a case for the proposed algorithm. Use larger matrices (randomly generated or from real data sets) with more entries missing. In fact matrix completion deals with cases where most of the values are missing. For example, you can try numerical simulations similar to the ones in "A Gradient Descent Algorithm on the Grassman Manifold for Matrix Completion' by Keshavan and Oh, or "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise' by Wainright and Negahban.

The algorithm presented in the paper has $(m-r)m$ optimization variables, so that it is suited for problems with small m and $m-r$ (number of rows and rank reduction). In recommender system applications, both dimensions of the matrix are big and the rank reduction is big. (In this application the rank is typically small.) This setup is better suited for the alternating projections type methods, which use image representation ($r(m+n)$ optimization variables). In addition, special care is needed in order to exploit the sparsity of the data matrix (in real-life recommender system applications most elements are missing). None of these issues were taken under consideration in our work.

As implemented in

> I. Markovsky and K. Usevich. Structured low-rank approximation with missing values. Technical Report 340718, Univ. of Southampton, `http://eprints.soton.ac.uk/340718`, 2012

the algorithm in the paper is applicable only to small size matrices, say, m up to 10 and n up to 100, so that it is of academic interest only. The efficient C implementation of

> I. Markovsky and K. Usevich. Software for weighted structured low-rank approximation. Technical Report 339974, Univ. of Southampton, `http://eprints.soton.ac.uk/339974`, 2012

however, can solve mosaic-Hankel-like low-rank approximation problems (an unstructured matrix is a special case of mosaic-Hankel matrix) with, say, m up to 100 and n up to 10^4.

In view of the above explanation, we have revised the numerical examples section with large problems: 10×100 and 10×1000 matrices of rank 9 with randomly distributed missing values. Going to more realistic recommender system problems requires modification of the method and the related software, which would make it problem dependent. Although this is possible, it would shift the focus of the paper outside the scope that we intend to address.

1.11: Also there are many matrix completion algorithms that works better than singular value thresholding. It would be interesting to compare the performance to LMaFit algorithm in "Solving A Low-Rank Factorization Model for Matrix Completion by A Nonlinear Successive Over-Relaxation Algorithm' by Zaiwen Wen, Wotao Yin, and Yin Zhang. There is a Matlab implementation which is powerful, public and extremely easy to use. (http://lmafit.blogs.rice.edu/) You might also want to try OptSpace algorithm from "matrix completion from a few entries' by Keshavan, Montanari, Oh (http://www.stanford.edu/ raghuram/optspace).

We have revised to the numerical examples in the section "Unstructured matrix with missing data" and, as suggested, compare now the methods in the paper with four alternative methods.

In the case of unstructured data matrix the methods in the paper are compared with the following alternative methods:

`wlra` — the alternating projections method of [Mar11],

`optspace` — a method based on spectral techniques and manifold optimization [KMO10],

`lmafit` — the successive over-relaxation algorithm of [WYZ10], and

`rtrmc` — the Riemannian trust-region method of [BA11].

Although, the examples are still small compared with the typical examples used in the recommender systems literature, they are challenging because of existence of local minima. In all experiments, our methods compute identical results with `rtrmc`. The LMaFit algorithm is the fastest of all compared methods.

1.12: For system identification, it is not clear how easy or difficult this example is. Add a result on the same problem using other approaches for solving this problem, perhaps using nuclear norm minimization approaches.

System identification with missing data is a research topic with its own literature, see, *e.g.*,

> A. Isaksson. Identification of ARX-models subject to missing data. *IEEE Trans. Automat. Control*, 38(5):813–819, May 1993
>
> R. Pintelon and J. Schoukens. Frequency domain system identification with missing data. *IEEE Trans. Automat. Control*, 45(2):364–369, February 2000

In our experience the nuclear norm heuristic is not effective in solving system identification problems, see

> I. Markovsky. How effective is the nuclear norm heuristic in solving data approximation problems? In *Proc. of the 16th IFAC Symposium on System Identification*, Brussels, 2012

We include comparison with the method of

> R. Pintelon and J. Schoukens. Frequency domain system identification with missing data. *IEEE Trans. Automat. Control*, 45(2):364–369, February 2000

(Software implementation was kindly provided to us by the authors.) This method is also based on local optimization and the simulation results show that its performance is close to the one of the proposed methods.

1.13: Also for data-driven simulations, compare the performance of the proposed algorithm to other known algorithms for this problem.

The only alternative method we are aware of is the subspace method of

> I. Markovsky and P. Rapisarda. Data-driven simulation and control. *Int. J. Control*, 81(12):1946–1959, 2008

In the revised version of the paper, we include a comparison with this method. Subspace methods are multi stage methods, *i.e.*, they split the non-convex optimization problem in several steps that are individually convex, but do not guarantee global or local optimally of the overall problem. As expected, the subspace method is more efficient but less accurate than the local optimization method used in the paper.

1.14: 1. In the abstract a citation should be changed to [13].

The citation is now removed (see the answer to question 1.1).

1.15: 2. In page 3, n_m is used with out properly being defined.

n_m is the number of missing elements. It is now defined in subsection "Notation".

$\mathscr{M}$ / $\mathscr{G}$ is the vector of indices of p (in decreasing order) that are missing / given, and n_{m} / n_{g} is the number of missing / given elements.

1.16: 3. In page 5, use $W_g^{1/2}$ instead of $\sqrt{W_g}$.

Done

1.17: 4. In page 5, The problem (WSLRA') is not well defined since not all values of p are given. How is the objective function defined when we have missing values?

We use the `NaN` symbol as a place holder and do not defined how arithmetic operations are carried out with it. This is not needed in the paper, except for Note 6. We agree with the reviewer that a careful analysis of the general weighted structured low-rank problem problem (WSLRA') is needed and have removed Note 6.

1.18: 5. In page 5, in note 7, how can W have dimensions mn*mn, when p has dimension at most n_p? Because of this, the rest of note 7 does not make sense.

For an unstructured matrix $n_p = mn$ (all elements of the matrix are parameters). In note 7, the weight matrix, refers to a unstructured low-rank approximation problem, so that the number of parameters in this problem (not the original structured one) is nm. The approximation matrix $\widehat{D}$ in the unstructured problem however turns out to be structured. This structure is not imposed explicitly as a hard constraint, but occurs implicitly by the weighted cost function of the optimization problem.

Note 1 (Solving (SLRA) as weighted unstructured problem). Consider an instance of problem (SLRA), refered to as problem P1, with structure $\mathscr{S} = \mathscr{S}_1$ and an instance of problem (WSLRA), refer to as problem P2, with unstructured correction ($\mathscr{S}_2 = \mathrm{vec}^{-1}$, $n_{p_2} = mn$) and weight matrix

$$W_2^{-1} = \mathbf{S_1}\mathbf{S_1}^\top. \qquad (\mathscr{S}_1 \mapsto W_2)$$

It can be verified by inspection that the cost functions (M) and (M_W) of problems P1 and P2, respectively, coincide. The weight matrix $W_2 \in \mathbb{R}^{mn\times mn}$, defined in ($\mathscr{S}_1 \mapsto W_2$), however is singular ($\mathrm{rank}(W_2)$ is equal to the number of structure parameters of problem P1, which is less than mn). In the derivation of the cost function (M_W) it is assumed that W_{g} is positive definite, so that minimization of (M_W) is not equivalent to problem P2.

1.19: 7. In page 7, note 9, This particular initialization of using SVD on zero-filled matrices has been rigorously analyzed in the context of matrix completion in "Matrix completion from a few entries' by Keshavan, Montanari and Oh (Theorem 1.1).

Thank you for the comment and the reference. The reference has been included in the paper to support the choice of the initial approximation by filling in zeros for the missing values.

Answer to referee #2

2.1: Section 4, It is a little bit strange to refer to [14] for "general purpose constrained optimization methods" and then propose "another approach" based on quadratic penalty, which is one of the methods presented in [14]. (In fact, many methods in [14] are penalty methods.)

We now refer to the second optimization method in the paper as a penalty method. We believe, however, that it should be distinguished from the formulation of the problem that uses the quadratic equality constraint $RR^\top = I$.

2.2: I believe the authors should mention (and investigate!?) the optimization methods on manifold, see "Optimization Algorithms on Matrix Manifolds, P.-A. Absil, R. Mahony, and R. Sepulchre, Princeton University Press".

We added the following explanation for the link to optimization on a manifold:

(SLRA$'_R$) is an optimization problem on a Stiefel manifold [AMS08] and can be solved by specialized methods, *e.g.*, the GenRTR package [BAG], or by general purpose penalty methods for constrained optimization [NW99].

2.3: Note that there is a Matlab toolbox available: `http://www.math.fsu.edu/~cbaker/GenRTR/`. It can be used, for example, for low-rank matrix completion; see, e.g., "N. Boumal and P.-A. Absil, RTRMC:

Thank you for the comment and the reference. We are currently using the RTRMC package in an extension paper of

> 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

2.4: Maybe authors should also motivate the fact that they use local optimization techniques (the problem is NP-hard in general since LRA with missing data is).

We added the following motivation:

Both the structured low-rank approximation and approximate matrix completion problems are nonconvex optimization problems that in general admit no analytic solution. Therefore, in both domains local optimization and convex relaxation heuristics are used as solution techniques. In this paper, we use the local optimization approach.

2.5: Section 5.1, The comparison is not very convincing. Authors should address the following questions: What is the computational complexity of their algorithm? It seems that, since $R \in \mathbb{R}^{(m-r)\times r}$, it will be rather expensive, especially if m is large.

Detailed analysis of the computational complexity of the general algorithm for affine structured weighted low-rank approximation and the specialized ones for mosaic-Hankel-like low-rank approximation problems is done in

> K. Usevich and I. Markovsky. Variable projection for affinely structured low-rank approximation in weighted 2-norm, 2012. Available from `http://arxiv.org/abs/1211.3938`

We have included the following paragraph in the revised version of the paper, addressing the question about the computational complexity.

The general solution method proposed in the paper has computational complexity $O(n^3)$ per iteration, where n is the number of columns of the data matrix. This makes it unsuitable for large scale applications. In special cases, such as unstructured matrix with missing data and Hankel structured matrix, there are efficient $O(n)$ methods. Modification of the methods in the paper for efficient computation in the case of mosaic-Hankel [Hei95] matrices, is possible and will be presented elsewhere.

2.6: What are the limits for m, n and r?

They depends on the implementation of the method and the type of structure. As implemented in

I. Markovsky and K. Usevich. Structured low-rank approximation with missing values. Technical Report 340718, Univ. of Southampton, `http://eprints.soton.ac.uk/340718`, 2012

the algorithm in the paper is applicable only to small size matrices, say, m up to 10, n up to 100, and rank reduction $m - r$ up to 5. The efficient C implementation

I. Markovsky and K. Usevich. Software for weighted structured low-rank approximation. Technical Report 339974, Univ. of Southampton, `http://eprints.soton.ac.uk/339974`, 2012

of the method, however, can solve mosaic-Hankel-like low-rank approximation problems with, say, m up to 100, n up to 10^4, and rank reduction up to 10.

We have added the following explanation in the paper:

Section 3:

Note 2 (Efficient computation and software implementation). Efficient evaluation of the cost function and its derivatives in the special case of mosaic-Hankel matrix structure *without* missing values is presented in [UM12b]. The method, presented in this paper (general linear structure) and the efficient methods of [UM12b] are implemented in Matlab (using Optimization Toolbox) and in C++ (using the GNU Scientific Library [gsl] for the optimization methods), respectively. Description of the software and overview of its applications is given in [MU12a].

Section 5.1:

The implementation [MU12b] of the methods in the paper is applicable only for small size problems (say, $m < 10$, $n < 100$, and $m - r < 3$). For larger problems, the efficient C implementation [MU12a] (denoted by [[slra-c]] below), of the variable projection approach can be used by setting small values for the weights corresponding to the missing values.

2.7: I believe it is nice to show that the method can be applied on the missing data problem on a toy example (a 3-by-10 matrix to be approximated with a rank-one matrix). However, as the proposed method is rather general, I would be very surprised if it competes with methods designed exclusively for that problem (e.g., the method referred above). Authors should comment on that (because it is not very clear in the text). In fact, authors seem to conclude that their method is competitive. Is that really the case?

We now state clearly that the objective of the paper is the development of a general algorithm which is not competitive to the existing state-of-the-art methods when specialized to cases, such as unstructured matrix completion.

Introduction:

The general solution method proposed in the paper has computational complexity $O(n^3)$ per iteration, where n is the number of columns of the data matrix. This makes it unsuitable for large scale applications. In special cases, such as unstructured matrix with missing data and Hankel structured matrix, there are efficient $O(n)$ methods. Modification of the methods in the paper for efficient computation in the case of mosaic-Hankel [Hei95] matrices, is possible and will be presented elsewhere.

Conclusion:

The performance of the methods in the paper was illustrated on small-size simulation examples and was compared with the performance of problem specific methods. Efficient computation for large scale problems appearing in applications such as recommender systems and system identification is a topic of future research.

2.8: Authors mention "Number of iterations for convergence": convergence in what sense? (objective function? Norm of the gradient? Difference between iterates?)

In the revised version of the paper we do not compare the number of iterations for convergence but the relative approximation error and execution time. (Convergence is checked with respect to different criteria, some of which are not used by all methods, so that we can not ensure that all methods stop under the same conditions.)

2.9: (p.3) Replace $p \in \mathbb{R}^{n_p} \cup \texttt{NaN}$ with $p \in (\mathbb{R} \cup \texttt{NaN})^{n_p}$?

Done

2.10: (p.3) I would rather write something like: "..., evaluate the cost function $M(R)$, ..., and find a point p that attains the minimum $M(R)$.

The problem statement is revised as suggested.

2.11: (p.3) "has a unique minimum $M(R)$" should be changed: a minimum objective function value is always unique (if it exists). I would suggest something like: Problem 1 has a unique global minimum—put equation (?) here—with objective function value—put equation (M) here.

The theorem statement is revised as suggested.

2.12: (p.3) n_m is not defined.

n_m is the number of missing elements. It is now defined in Subsection "Notation" of the introduction.

$\mathscr{M}$ / $\mathscr{G}$ is the vector of indices of p (in decreasing order) that are missing / given, and n_{m} / n_{g} is the number of missing / given elements.

2.13: (p.5) I think it would be better to put Section 3 after Section 4. The paper reads very well until then. My opinion is that it would be easier for the reader (at least it would have been for me) to go on with Section 4 after Section 2, and later on explain how it can be generalized. That is, follow: (1) Problem formulation, (2) Inner problem, (3) Outer problem, (4) Generalization.

Done

2.14: (p.6) I really like the result that there exists $\gamma = \max_R M(R)$ so that the two problems are equivalent (this is in general not the case [14] — for quadratic penalty, usually only convergence for $\gamma \to +\infty$ is guaranteed). However, it would be nice to mention how γ is evaluated in practice.

In the numerical examples, we take the value $\gamma = \|p_{\mathscr{G}}\|_2^2$, see Note 7 in the revised version of the paper. Choice of γ is now explicitly stated in Section 5.

2.15: (p.7) Note 9. Why would you fill in the missing entries with zeros? Wouldn't it make more sense to use the average of the non-missing entries? Or the average by row/column?

As pointed out by Review 1, justification for the zero imputation heuristic is given in Theorem 1.1 of

> R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, August 2010

2.16: (p.7) It solveS low-rank ...

Corrected

2.17: (p.9) Conclusion. "Two optimization strategies were proposed" ... In this text, only one is (briefly) described (the one using quadratic penalty) so writing "Two optimization strategies were proposed" seems a little bit abusive especially since optimization strategies in [14] are countless. In other words, I would not say that writing "... can be solved by general purpose constrained optimization methods" amounts to proposing an optimization strategy.

The conclusion was rewritten as follows:

A variable-projection-like approach for structured low-rank approximation with missing data was developed. The approach was furthermore generalized to weighted structured low-rank approximation with missing values. After elimination of the approximation $\widehat{p}$, the remaining nonlinear least-squares problem subject to quadratic equality constraints was solved as an equivalent regularized unconstrained optimization problem.

The problem and solution methods developed have applications in matrix completion (unstructured problems), system identification with missing data, and data-driven simulation and control (mosaic-Hankel structured problems). The performance of the methods in the paper was illustrated on small-size simulation examples and was compared with the performance of problem specific methods. Efficient computation for large scale problems appearing in applications such as recommender systems and system identification is a topic of future research.

2.18: By the way, it would be good to mention which methods are implemented in [10,11].

This is now mentioned in the note "Efficient computation and software implementation":

The method, presented in this paper (general affine structure) and the efficient methods of [UM12b] are implemented in Matlab (using Optimization Toolbox) and in C++ (using by the Levenberg-Marquardt algorithm [Mar63] from the GNU Scientific Library [gsl]), respectively.

References

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[BA11] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 406–414. 2011.

[BAG] C. Baker, P.-A. Absil, and K. Gallivan. GenRTR riemannian optimization package. `http://www.math.fsu.edu/~cbaker/GenRTR`.

[CP09] E. Candes and Y. Plan. Matrix completion with noise. *arXiv:0903.3131*, March 2009.

[CR09] E. Candés and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2009.

[GP03] 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.

[gsl] GSL — GNU scientific library. `www.gnu.org/software/gsl/`.

[Hei95] G. Heinig. Generalized inverses of Hankel and Toeplitz mosaic matrices. *Linear Algebra Appl.*, 216(0):43–59, February 1995.

[Isa93] A. Isaksson. Identification of ARX-models subject to missing data. *IEEE Trans. Automat. Control*, 38(5):813–819, May 1993.

[KMO10] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, August 2010.

[LHV12] Z. Liu, A. Hansson, and L. Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Submitted to System and Control Letters*, 2012.

[Lju] L. Ljung. *System Identification Toolbox: User's guide*. The MathWorks.

[Mar63] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

[Mar08] I. Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008.

[Mar11] I. Markovsky. *Algorithms and literate programs for weighted low-rank approximation with missing data*, volume 3 of *Springer Proc. Mathematics*, pages 255–273. Springer, 2011.

[Mar12a] I. Markovsky. How effective is the nuclear norm heuristic in solving data approximation problems? In *Proc. of the 16th IFAC Symposium on System Identification*, Brussels, 2012.

[Mar12b] I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer, 2012.

[MGSF97] B. De Moor, P. De Gersem, B. De Schutter, and W. Favoreel. DAISY: A database for identification of systems. *Journal A*, 38(3):4–5, 1997. Available from `http://homes.esat.kuleuven.be/~smc/daisy/`.

[MR08] I. Markovsky and P. Rapisarda. Data-driven simulation and control. *Int. J. Control*, 81(12):1946–1959, 2008.

[MU12a] I. Markovsky and K. Usevich. Software for weighted structured low-rank approximation. Technical Report 339974, Univ. of Southampton, `http://eprints.soton.ac.uk/339974`, 2012.

[MU12b] I. Markovsky and K. Usevich. Structured low-rank approximation with missing values. Technical Report 340718, Univ. of Southampton, `http://eprints.soton.ac.uk/340718`, 2012.

[MV78] C. Moler and Ch. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836, 1978.

[MV07] I. Markovsky and S. Van Huffel. Overview of total least squares methods. *Signal Proc.*, 87:2283–2302, 2007.

[MVP05] 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.

[MWRM05] I. Markovsky, J. C. Willems, P. Rapisarda, and B. De Moor. Algorithms for deterministic balanced subspace identification. *Automatica*, 41(5):755–766, 2005.

[NW99] J. Nocedal and S. Wright. *Numerical optimization*. Springer-Verlag, 1999.

[PS00] R. Pintelon and J. Schoukens. Frequency domain system identification with missing data. *IEEE Trans. Automat. Control*, 45(2):364–369, February 2000.

[UM12a] 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.

[UM12b] K. Usevich and I. Markovsky. Variable projection for affinely structured low-rank approximation in weighted 2-norm, 2012. Available from `http://arxiv.org/abs/1211.3938`.

[WYZ10] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Technical report, Rice University, 2010. CAAM Technical Report TR10–07.