

Author's response to the referees' comments

Manuscript title: Structured low-rank approximation and its applications

Author: Ivan Markovsky

May 23, 2007

I thank the referees and the associate editor for their relevant and useful comments. In this document, I quote in **bold face** statements from the reports. My replies follow in ordinary print.

Major changes in the revised version of the manuscript

1. Section 5 “Algorithms”¹ is revised according to the comments of referee 2 (see the third comment on page 7). The following subsections are added:
 - “Special cases with known analytic solutions” — contains the material of Section 2 “Low-rank approximation and total least squares” in the original version and a result on global solution of circulant structured low-rank approximation problems,
 - “Suboptimal solution methods” — present methods based on SVD heuristics and convex relaxations, and
 - “Simulation example” — shows an example of DAISY [DM05] illustrating the application of the structured low-rank approximation method on real data of wing flutter.
2. As suggested by referee 2 (see the first comment on page 7), material from Section 4.1 “Errors-in-variables identification” is moved to Section 3.2 “Structured low-rank approximation”.
3. Section 4.2 “Harmonic retrieval” is revised. The mistake pointed out by the associate editor (see comment (6) on page 4) is corrected and the generality of the statements is restricted to the scalar case.
4. Section 6.1 “Weighted low-rank approximation” is extended with extra material — heuristic methods and local optimization methods for this class of problems.
5. Section 6.2 “Nonnegative low-rank approximation” is revised. The mistake pointed out by referee 1 (see the last comment on page 6) is corrected and the concept of a positive rank of a matrix is introduced.
6. According to the suggestion of referee 3 (see the first comment on page 11), Section 6.3 “Kernel mapping for a second order model fitting” is removed.

¹Section numbers refer to the original version of the manuscript.

Answer to the comments of the associate editor

(1) The main statement of the paper: 'Behind every data modeling problem there is a (hidden) low-rank approximation problem: ...' is somewhat overdone. (In addition the term 'linear' is missing from this statement!)

As shown in Section 6.3, fitting quadratic static models to data can also be posed as a low-rank approximation problems. For this reason I did not restrict the statement to the linear case. I agree, however, that the presentation is essentially limited to the linear case and (as pointed out by referee 3) the kernel mapping of Section 6.3 is not a "nice" way of addressing nonlinear modeling problems, so that the generality of the statement should be restricted to the linear case.

There is a wide range of modeling approaches, which do not use such an interpretation. E.g. modeling and estimation of AR(MA) processes do not fall into this scheme. (At least, it would be an artificial interpretation.)

For some of these approaches the reason may be that nobody yet looked at the corresponding problems from the low-rank approximation point of view. Preliminary results showing application of low-rank approximation to ARMAX system identification are reported in [MWD06b].

Despite my own conviction in the generality of the statement, I agree that it is currently more a believe than a well established fact, so I rephrased is as follows:

It is the believe of the author that behind every linear data modeling problem there is a (hidden) low-rank approximation problem: the model imposes relations on the data which render a matrix constructed from exact data rank deficient. Low-rank approximation is used in a number of data modeling problems from diverse scientific fields, however, there are problems, e.g., frequency domain and stochastic system identification, that are still awaiting for such an interpretation.

(2) One should clearly differentiate between the representation of a line (e.g. $b = ax$, $a = bx$, $(a, b)' = Pl$, $R(a, b)' = 0$) and the 'miss fit' criterion (e.g. $(a - \hat{a})^2$, $(b - \hat{b})^2$, $(a - \hat{a})^2 + (b - \hat{b})^2$, ...). The difference of the two fitting lines presented in Fig. 1 are due to the use of two different fitting criterion (and not due two different representations)!

I completely agree. The reason for the confusion (and this is indeed a major issue illustrated by the example) is that the "classical method" for data fitting (i.e., solve approximately a linear system of equations by the least squares method) do not differentiate the model representation and the fitting criterion. In the classical approach, one can not choose the model representation and the fitting criterion independently. For example, solving $b = ax$ in the least squares sense automatically minimizes the sum of the squared vertical distances from the data points to the fitting line.

The framed sentence: 'the least squares fit to the data is affected by the choice of the model representation.' is not true. It should be something like 'the optimal fitted line is affected by the choice of the miss fit criterion.'

To be precise and avoid the confusion "least squares fit" should have been written explicitly as "solving approximately a system of equations by the least squares method". The point is that in the classical method the fitting criterion is bound to the model representation. The framed sentence is revised as follows:

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

Of course the representations $b = ax$ and $a = bx$ resp. do not cover all lines in \mathbb{R}^2 . Therefore it may happen that there is no solution, since the optimal line has no representation of the given form.

The consequences of the $b = ax$ and $a = bx$ representations:

- *nongeneric problems*, leading to theoretical complications, see [VV91],

- *ill-conditioning*, leading to algorithms that lack numerical robustness, and
- *need of regularization*, leading to a change of the specified fitting criterion

are deemed inherent to the data fitting problem per se and are accepted as a necessary (obligatory?) aspect of data-modeling. The example shows that these complications are inherent in the classical setting (using a fixed input/output representation) and not in the original data fitting problem. Obviously they can be avoided by choosing the alternative image and kernel model representations. An explanation is added in the end of Section 1.2 to make this point clear.

(3) ... the definitions are sloppy and incomplete. For the static case the definition should be: A linear! static model \mathcal{B} with d variables is a linear! subset of \mathbb{R}^d .

First, I give the general definition of a static model (a subset of the data space). Then, I specialize it to the linear case (a subspace of the data space). However, as stated, the general definition is rather confusing: it appears without warning right after the heading “Linear static model” and before the list of representations for linear static models. In addition the definition of a linear static model was missing. The passage is revised as follows:

A static model \mathcal{B} with d variables is a subset of \mathbb{R}^d . A linear static model with d variables is a subspace of \mathbb{R}^d . Three basic representations of a linear static model $\mathcal{B} \subseteq \mathbb{R}^d$ are ...

For the dynamic case the subset \mathcal{B} has to be linear! and closed!

Again, first, I give the general definition and, then, I specialize it to the linear case. Note that “ \mathcal{B} is a linear subset” is implied by “ \mathcal{B} is a subspace”. The condition “ \mathcal{B} closed” corresponds to the property that “ \mathcal{B} is finite-dimensional”. The passage is revised as follows:

A discrete-time dynamic model \mathcal{B} with w variables is a subset of $(\mathbb{R}^w)^\mathbb{N}$, the set of all functions from the variable space \mathbb{R}^w to the time axis $\mathbb{N} := \{1, 2, \dots\}$. By definition, a model \mathcal{B} is LTI if \mathcal{B} is a shift-invariant subspace of $(\mathbb{R}^w)^\mathbb{N}$. If in addition, \mathcal{B} is a closed subset (in the topology of point-wise convergence), then it turns out that \mathcal{B} is finite-dimensional.

In addition the terms ‘order’ and ‘lag’ need to be defined and explained. A reference to a more detailed analysis should be given.

The missing definitions are added. References are cited in the main text where the reader is directed to the appendix:

Appendix 1 collects basic facts, used in the paper, about LTI models and their representations, see also [PW98,MWVD06].

(4) Proposition 8: I assume that the data w_d is finite. Then, like in the other examples, the problem formulation should be $w_d = \bar{w} + \tilde{w}$, where $\bar{w} \in \tilde{\mathcal{B}}_T$, $\tilde{\mathcal{B}} \in \mathcal{L}_{m,\ell}^w$ and ...

The missing restriction of \mathcal{B} on the interval $[1, T]$ is added.

Furthermore the models \mathcal{B} are linear subsets of $(\mathbb{R}^w)^\mathbb{N}$, thus it does not make sense to say that the estimator $\hat{\mathcal{B}}^*$ is asymptotically normal. (Of course, if one uses a parametric representation of the models, then the estimated parameters may be asymptotically normal.)

The text is corrected as follows:

Under additional stochastic assumptions, see [PS01,KMV05], the estimator $\hat{\mathcal{B}}^*$ is consistent and the estimated parameters have asymptotically normal distribution.

(5) Section 4.3 Model reduction: It should be clearly indicated that the impulse response is based on an input/output representation. Thus the proposed norm for models is not representation free, but depends on a chosen input/output decomposition of the variables.

I agree. The comment is relevant for Sections 4.2-4.5 and 4.7 as well. The text is modified and the following explanation is added before Section 4.1.

Most of the work on the errors-in-variables identification problem (see [Söd07] and the references there in) is presented in the classical input/output setting, i.e., the proposed methods aim to derive a transfer function, matrix fraction description, or input/state/output representation of the system. The salient feature of the errors-in-variables problems, however, is that all variables are treated on an equal footing as noise corrupted. Therefore the input/output partitioning implied by the classical model representations is irrelevant in this problem. Section 3.1 presents the errors-in-variables identification problem in a representation free setting.

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

(6) Section 4.8 Harmonic retrieval: Not every marginally stable model \mathcal{B} has a palindromic kernel representation. E.g., consider the case $R(z) = 1 - z$.

The marginal stability constraint in harmonic retrieval problems is often relaxed to the easier to deal with constraint that the filter coefficients are palindromic. Looking in the literature, however, I was not able to find a theoretical justification for this. In the univariate case, it turns out that every marginally stable model \mathcal{B} has either a palindromic or antipalindromic ($R_i = -R_{\ell-i}$, for all i) kernel representation, see

<http://www.ecs.soton.ac.uk/~im/palindromic.pdf>.

The antipalindromic case occurs only when the system has a pole at 1 with odd multiplicity, so that the palindromic case is generic in the set of the marginally stable systems.

I've restricted the statements in Section 4.8 to the scalar case and revised the text.

Of course Problem 22 is a very 'strong' relaxation of Problem 21.

My experience is that it is a “weak” relaxation in the sense that if the data is a noisy trajectory of a marginally stable system, Problem 22 produces a marginally stable solution unless the SNR is very low (so low that the estimated parameters become meaningless).

In my opinion the proof of Proposition 23 is not complete. It shows that $\mathcal{S}(p)$ is rank deficient for models with a palindromic kernel. But I can't see why the reverse direction holds.

The reverse direction holds if the structure is $\mathcal{S}(p) = [\mathcal{H}_{\ell+1}(p) \quad \mathcal{T}_{\ell+1}(p)]$ instead of $\mathcal{S}(p) = \mathcal{H}_{\ell+1}(p) + \mathcal{T}_{\ell+1}(p)$. This is corrected.

(7) Section 4.9 Approximate divisor: I think that the size of the Sylvester structured matrix should be $n - d + 1$!

Yes, the size of the Sylvester structured matrix should be $n - d + 1$. This is corrected.

(8) variable projection method: The mapping G should be constructed such that $\text{vec}(R\mathcal{S}(p)) = G(R)p$ holds.

Corrected, see (21).

Answer to referee #1

The statements concerning the contribution of the paper in the introduction are a bit confusing. For example, the first sentence in the second paragraph of the paper reads "The purpose of the paper is to show that the abstract problem of fitting linear models to data is actually equivalent to approximating the matrix $[A \ B]$ by a low-rank matrix". However, after reading the paper I am not convinced that this is indeed the purpose of the paper. As written at the end of Section 1.3, the paper has several purposes, among them is to give a review of applications, present generalizations and give a "unifying viewpoint of a representative set of data modelling problems". Therefore, a more precise statement is in order.

The statement is revised as follows:

In this paper we show that a number of linear data fitting problems are equivalent to the abstract problem of approximating a matrix D constructed from the data by a low-rank matrix.

The "unifying viewpoint" mentioned in the end of Section 1.3 is the low-rank approximation problem, so that the statement of the introduction is correct and is consistent with the precise objectives of the manuscript. At the early stage of the exposition, I prefer the statements to be concise although this may make them less precise.

"Fitting models to data can be achieved ...". It will be more accurate to write "Fitting linear models to data can be achieved ...".

Corrected.

Beginning of Section 1.1: "and solve the undetermined system ...". This is not an accurate statement since essentially you solve this system approximately.

The sentence is corrected as follows:

The classical approach for line fitting is to define $\text{col}(a_i, b_i) := d_i$ and solve approximately the overdetermined system $\text{col}(a_1, \dots, a_N)x = \text{col}(b_1, \dots, b_N)$ by the least squares method.

Section 1.1. There is an inconsistency in the line fitting problem. Notice that \hat{D} is allowed to be a zero-matrix. Moreover, the conditions prior to the frame at the end of the section are not equivalent to the constraint $\text{rank}(\hat{D}) = 1$ but rather to the constraint $\text{rank}(\hat{D}) \leq 1$.

I agree—the constraint should be $\text{rank}(\hat{D}) \leq 1$.

Page 2, line 5. "achieving low-rank approximation" should be "achieving a low-rank approximation".

Corrected.

Page 2, line 12. "on a most simple data fitting problem ..." should be "on the most simple data fitting problem ...".

Corrected.

Page 4, line 11. $R \in \mathbb{R}^{2 \times 1}$ should be replaced with $R \in \mathbb{R}^{1 \times 2}$.

Corrected.

Page 5. The sentence inside the frame should be confined to linear models. It is not likely that every data modeling problem reduces to a low-rank approximation problem. In any case, the conclusions should be restricted to the topics discussed within the paper.

I agree. See the answer to the first two comments of the associate editor (page 2) for a detailed explanation.

Page 6. The second contribution should read "complete the list of applications ..." instead of "complete the listed of applications ...".

Corrected.

Page 9, line 7. "...is optimal approximate ..." should be "...is an optimal approximate ...".

Corrected.

Page 20, line 8. "As the structured ..." should be "Similarly to the structured ...".

Corrected.

Page 20, line -9. Why is it true that problem (25) is equivalent to the proceeding problem? It is obviously true that $\text{rank}(\hat{D}) \leq m$ if and only if $\hat{D} = PL, P \in \mathbb{R}^{d \times m}, L \in \mathbb{R}^{m \times N}$. Does it imply that $\text{rank}(\hat{D}) \leq m$ with $\hat{D}_{ij} \geq 0$ for every i, j if and only if $\hat{D} = PL, P \in \mathbb{R}^{d \times m}, L \in \mathbb{R}^{m \times N}$ with $P_{is}, L_{sj} \geq 0$?

The "if" part is trivially true but the "only if" is not correct. Problem 25 is actually a *relaxation* of the original problem. The minimal inner dimension m in the factorization $\hat{D} = PL, P \in \mathbb{R}^{d \times m}, L \in \mathbb{R}^{m \times N}$ with $P_{is}, L_{sj} \geq 0$ of an element-wise positive matrix \hat{D} (called "positive rank") is greater than or equal to the rank. For example the matrix

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

has rank 3, but does not have a nonnegative decomposition with inner dimension 3, so the positive rank is equal to 4.

The text is revised in order to make clear that the problem proceeding (25) is not equivalent to but a relaxation of (25).

Answer to referee #2

Structure and content: There is some repetition between sections 2 and 3.1, and between 3.1 and 4.1; these sections can be merged.

Section 2 presents approaches for solving (unstructured) low-rank approximation problems and is merged with Section 5 “Algorithms”. Proposition 8 of Section 4.1 is moved to Section 3.2. In this way, the presentation of Section 3.2 completely parallels the one of Section 3.1. (Proposition 8 is the dynamic equivalent to Proposition 6).

Content of section 1: In the contributions section, it is perhaps more accurate to say "the SLRA problem has not been surveyed as a data modeling principle" rather than "has not been proposed".

Corrected as follows:

Although omnipresent, however, by now the structured low-rank approximation problem is not generally perceived as a data modeling principle.

Except for the work of De Moor see, e.g., [DM93], structured low-rank approximation is presented in the literature exclusively as a numerical linear algebra problem. Its relevance to applications is often confined to a few statements in the introduction: there is no interpretation of the meaning of the structure and the rank constraint, no specific examples, no practical applications. In the context of the rank minimization problem Fazel [Faz02] also presents a comprehensive list of applications and gives insightful links. Apart from these two pieces of work, I am not aware of other references that would contradict to the word “generally” in the above statement. If the referee knows about work linking SLRA to data modeling, then I will appreciate knowing about it and will be glad to give it credit.

Note that in [Faz02] there is no reference to the work on the SLRA problem (in particular to the earlier work of De Moor) and vice versa, as far as I know, in the SLRA literature there is no reference to the rank minimization problem. Therefore, I believe the link between these two pieces of work is a contribution of the present paper.

Content of section 5: The algorithms section is incomplete, only discussing two local optimization approaches.

I agree. This section was restricted to the variable projections and alternating projections approaches that are best represented in the SLRA literature. A major revision step in the manuscript is to broaden the exposition to other approaches for solving the SLRA problem.

What categories of approaches are possible?

The general SLRA problem is an NP-complete optimization problem. Section 4.1 of the revised version of the manuscript, reviews two special cases that have analytic solutions: a weighted unstructured low-rank approximation problem and a circulant structured low-rank approximation problem. The general case, can be approached by heuristic methods based on convex relaxations (this approach is revised in Section 4.2), local optimization methods (Section 4.3), or global optimization methods (mentioned in the end of Section 4.2).

Are there any convex relaxations that can give a bound on the optimal value?

The obvious and most often used relaxation that leads to tractable problems is obtained by dropping the structure constraint. In the broader context of the rank minimization problem, convex relaxations are proposed in [Faz02].

Or, can branch and bound type algorithms be used for finding the global optimum?

Any global optimization method that can treat quadratic cost function and bilinear equality constraints can in principle be used for solving affine SLRA problems. For “small” problems global optimization is certainly an option. For other but “small” problems, however, global optimization is likely to be computationally too expensive. The approach based on global optimization is mentioned in the revised version of the manuscript but (because of the above reason) it is not revised to the same extent as the other approaches.

Alternating projections can be presented in a more general setting than the deconvolution problem. For example, can projections onto the set of matrices of a fixed rank be used? This is mentioned in section 6.1, but I think it would fit better in the algorithms section, and needs better discussion.

Projecting alternatively onto the set of unstructured matrices of a specified rank and onto the set of matrices with a specified structure is a well known idea. The algorithm is proposed by Cadzow in [Cad88], however, as proven in [DM94], it is not locally convergent, i.e., this algorithm converges to points that are not locally optimal for the SLRA problem. For this reason I did not include the algorithm of Cadzow in the overview paper. (It is a negative example for application of the alternating projections method.)

This approach is used for low-order controller design (see Skelton, Iwasaki and Grigoriadis, 1998) and can be relevant to problems discussed here as well.

A link between the alternating projections algorithm of Skelton, Iwasaki and Grigoriadis and the structured low-rank approximation problem is far from obvious and as far as I know is not reported in the literature. It may be the case that this algorithm solves an SLRA problem, however, it is a research question to find out which one.

Also algorithm 2 can be moved to section 5

This suggestion will lead to inconsistency, because Algorithm 2 will appear in a section about algorithms for the SLRA problem, however, it treats another problem.

Section 6.2 can be deleted since it a small variation on SLRA.

Section 6.2 illustrates the use of additional constraints. Nonnegative low-rank approximation is an active research topic in the machine learning community and has numerous applications. I do believe that it is useful to show this problem in the overview of structured low-rank approximation. Although it is a relatively small variation on SLRA it makes contact with a whole body of literature that is typically not considered as a part of systems and control.

How would the different local methods compare, in terms of the solution they converge to, and basic convergence properties? Can anything be said theoretically?

The variable projections method using the Levenberg–Marquardt algorithm is globally convergent with a superlinear convergence rate. The alternating projections algorithm is also globally convergent, however, with a linear local convergence rate. The theoretical properties of these methods suggest that variable projections, combined with a super linearly convergent algorithm for the nonlinear least squares problem, is more efficient than alternating projections. An explanation is added in Section 4.3.

If not, how do they compare in practice? If there are references that answer these questions, they can be added.

Several system identification methods (including the SLRA method of [MVP05]) are compared, see [MWD06a], on benchmark problems of the data base for system identification DAISY [DM05]. The general conclusion is that the heuristic methods, known as subspace identification methods, are faster but less accurate than the methods based on local optimization, such as the prediction error methods and the method of [MVP05]. It is a common practice to use a suboptimal solution obtained by a subspace method as an initial approximation for an optimization based method. The reference [DM05] and the above explanation are added in Section 4.2.

The log-det heuristics of [Faz02] and the heuristic obtained by dropping the structure constraint (Kung’s algorithm) are compared empirically on the approximate realization problem. The trade-off curves obtained by the two methods are shown on Figure 1. The result suggests that Kung’s algorithm is a more efficient heuristic for approximate realization than the log-det heuristic. The simulation files (using MATLAB and the LMI control toolbox) are available at:

`http://www.ecs.soton.ac.uk/~im/logdet_heuristic.tar` (UNIX tar file)
`http://www.ecs.soton.ac.uk/~im/logdet_heuristic.zip` (Windows zip file)

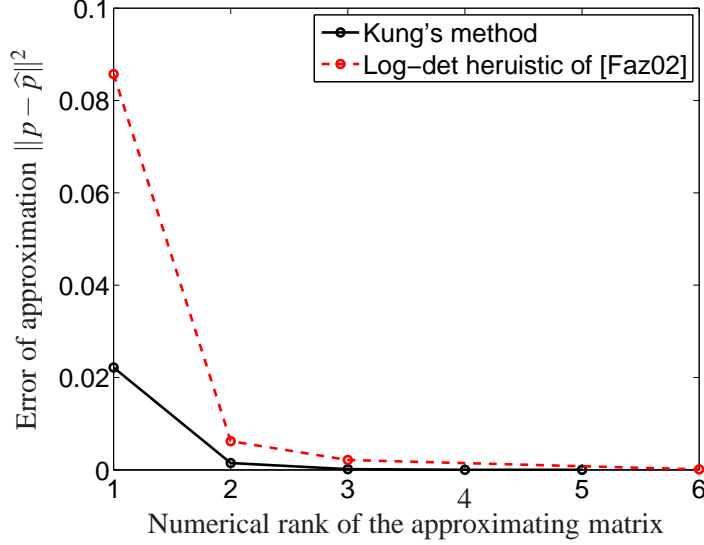


Figure 1: Trade-off curves (approximation error vs rank) for the log-det heuristic of [Faz02] (dashed) and Kung's method (solid), compared on a Hankel low-rank approximation problem. (Rank is computed with a tolerance e^{-8} .)

Content of section 6: Some related problems listed here are small variations of the SLRA (e.g., 6.2), and some are much broader categories (e.g., rank minimization problem)—perhaps the organization can be made more even.

An overview of heuristic methods and methods based on local optimization for the weighted low-rank approximation problem is added in Section 6.1. Section 6.2 is revised and extended, see page 1 for a detailed explanation on the major revision steps in the structure of the paper. The generality of the rank minimization problem (due to the general convex constraint) is emphasized, see my answer to the next comment.

The rank minimization problem is much more general than SLRA, and the relaxation and heuristic that are referred to ([Faz02]) can handle other convex constraints in addition to the error bound. This reference also includes a frequency domain low-order system approximation problem, which can also be mentioned.

The generality of the rank minimization problem is made clear in the revised version of the manuscript:

In [Faz02], the following formulation of the rank minimization problem is considered

$$\min_{\hat{D}} \text{rank}(\hat{D}) \quad \text{subject to} \quad \hat{D} \in \mathcal{C}, \quad (\text{RMP})$$

where \mathcal{C} is a convex set. Note that the convex constraint $\hat{D} \in \mathcal{C}$ is much more general than the norm constraint $\|p - \hat{p}\| \leq \gamma$ in (27).

The system approximation problem is mentioned in the end of Section 5.3:

A number of applications of the rank minimization problem are presented in [Faz02]. Among them are system realization with time domain constraints, reduced order controller design, and frequency domain system approximation.

For section 6.5, I think the structured distance to singularity also arises in condition number analysis of structured linear programs, this can be looked into and referenced if appropriate.

Condition number analysis of structured linear system of equations, see [Rum03], is relevant and is referred to in Section 5.4 of the revised version of the manuscript:

Closely related to the structured pseudospectra problem is the structured condition number problem for a linear system of equations, see [Rum03].

I would be glad to check specific references for condition number analysis of structured linear programs.

Writing and style: ... the sentences are occasionally confusing. For example, the last two sentences on page 5,

Revised as follows:

We consider the more general structured low-rank approximation problem because it gives freedom in choosing different representations for solving particular problems.

or second sentence on page 6 ...

Revised as follows:

Therefore, the question of representing the rank constraint in the low-rank approximation problem corresponds to the question of choosing the model representation in the data fitting problem.

Two grammar comments: the word 'whose' should replace 'which' in phrases such as 'a system which transfer function ...

Corrected.

it is better to replace phrases such as 'to-be-found model' with 'the model to be found'

The minor corrections are taken into account.

Answer to referee #3

1. p.20. Section 6.3 should be SKIPPED. Reason: algebraic fit leads to VERY poor estimator, which is inconsistent.

Done. Please see page 1 for a detailed explanation of the restructuring of the manuscript.

2. Everywhere in the paper there should be solution TO the problem or solution set TO the problem.

Corrected.

3. Letter l looks in formulae very much alike number 1. I recommend to choose another letter.

In the revised version of the manuscript, the lag of the system is denoted by ℓ .

4. p.4, after (8) ... have (possibly nonunique) solutions.

Corrected.

5. p.5, -12 l. dot at the end of the row.

Corrected.

6. p.8 in the middle ... about linear time-invariant (LTI) models ...

The term “linear time-invariant” appears first in Section 1.3. The definition of the acronym LTI is given there.

7. p.9: Say immediately after (10) that W is a positive semidefinite matrix, and v is a positive scalar.

Done. (Note that we require W to be positive definite. Otherwise $\text{vec}(\tilde{D})$ has a deterministic component.)

In Proposition 9: Then a measurable solution ... (because statistical estimator is a measurable mapping acting from a probability space)

Corrected.

-10 l. The right-hand side should be taken to the power $1/2$.

Corrected.

-2 l. ... leads to low-rank ...

Corrected.

8. p. 10, 5 l. Consider an affine mapping ... (other mappings are not considered here; Wandermond nonlinear structure leads to inconsistent estimators, and corresponding STLS problem should not be recommended)

The following note is added:

Nonlinearly structured matrices are not considered in this paper because the corresponding nonlinearly structured low-rank approximation problems are much harder to solve than the affine ones and in the errors-in-variables setting the corresponding maximum likelihood estimators are inconsistent.

-4 l. Notation for the interval is not good. Better: $1, T$ and bar above $1, T$.

The notation is changed as suggested.

9. p.11 -8 l. delete comma after section

Done.

10. p.12 -12 l. ... the true parameters are guaranteed ...

Corrected.

-10 l. ... optimal (asymptotic efficiency) ... (in errors-in-variables setting Cramer-Rao inequality does not work since the MLE does not have expectation; but the MLE in linear models has the smallest possible asymptotic covariance matrix)

Clarified as suggested:

... optimal (asymptotic efficiency) under certain specified conditions. In the errors-in-variables setting the maximum likelihood estimator does not have expectation, however, for linear models it has the smallest possible asymptotic covariance matrix.

11. p.13 -5 l. ... in the 2-norm sense.

Corrected.

12. p.14 common notation is deg for degree.

The notation "degree" is changed to "deg".

I think that the ratio of two polynomials is proper if $\deg P < \deg Q$, equality is not allowed. If the author wants to allow the equality then the term "proper" is not appropriate.

The text is rephrased in order to avoid the term proper.

13. P.18. Formula (21): reference to the paper is needed where this was obtained. And in (19), (20) make an interval after R , before RR^T .

Corrected.

14. P.22 -10 l. ... trade-off ... (because below this spelling will be used)

Corrected.

-4 l. ... solve another problem ...

Corrected.

15. P.28 -5 l. Remark. An image ... (because it is not a part of proof) The same on p.29, 1 line: Remark after: skipped.

Done.

16. p.15 in the middle no comma after Therefore, the same twice on p.30.

Corrected.

17. p.20 in (25) and the next formula: explain notation for the entries of matrices.

Above Section 6.2: ... work of Manton [MMH03] ... rank). There a kernel representation ...

Rephrased as suggested.

References

- [Cad88] J. Cadzow. Signal enhancement—A composite property mapping algorithm. *IEEE Trans. Signal Process.*, 36:49–62, 1988.
- [DM93] B. De Moor. Structured total least squares and L_2 approximation problems. *Linear Algebra Appl.*, 188–189:163–207, 1993.
- [DM94] B. De Moor. Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Trans. Signal Process.*, 42(11):3104–3113, 1994.
- [DM05] B. De Moor. DaISy: Database for the identification of systems. Dept. EE, K.U.Leuven, www.esat.kuleuven.be/sista/daisy/, 2005.
- [Faz02] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Elec. Eng. Dept, Stanford University, 2002.
- [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.
- [MWD06a] I. Markovsky, J. C. Willems, and B. De Moor. Comparison of identification algorithms on the database for system identification DAISY. In *Proceedings of the 17th Symposium on Mathematical Theory of Networks and Systems*, pages 2858–2869, Kyoto, Japan, 2006.
- [MWD06b] I. Markovsky, J. C. Willems, and B. De Moor. The module structure of ARMAX systems. In *Proceedings of the 41st Conference on Decision and Control*, pages 811–816, San Diego, USA, 2006.
- [Rum03] S. Rump. Structured perturbations part i: Normwise distances. *SIAM J. Matrix Anal. Appl.*, 25:1–30, 2003.
- [Söd07] T. Söderström. Errors-in-variables methods in system identification. *Automatica*, 43:939–958, 2007.
- [VV91] S. Van Huffel and J. Vandewalle. *The total least squares problem: Computational aspects and analysis*. SIAM, Philadelphia, 1991.