Authors' response to the referees' reports

Paper title: Application of Structured Total Least Squares for System Identification and Model Reduction

Authors: Ivan Markovsky, Jan C. Willems, Sabine Van Huffel, Bart L.M. De Moor, and Rik Pintelon

We thank the Editors and the referees for their comments and corrections.

In this document, we quote in **bold face** statements from the reports. Our replies follow in ordinary print.

Review of the Editors

One thing that needs to be further clarified is the link between "error-in-variables" (EIV) and "exact-input" (EI) techniques. Be explicit in IV:A to explain what you mean. Also in IV:B the mismatch $w - \hat{w}$ will be the common "output error".

In the EIV case, the considered identification problem is

$$\hat{\mathscr{B}}_{\text{eiv}} := \arg \min_{\mathscr{B} \in \mathscr{L}_{m,l}} \left(\min_{\hat{w}} \|w - \hat{w}\|_{\ell_2}^2 \quad \text{subject to} \quad \hat{w} \in \mathscr{B} \right) \tag{EIV}$$

and in the EI case, it is

$$\hat{\mathscr{B}}_{\text{ei}} := \arg \min_{\mathscr{B} \in \mathscr{L}_{m,l}} \left(\min_{\hat{y}} \|y - \hat{y}\|_{\ell_2}^2 \quad \text{subject to} \quad \text{col}(u, \hat{y}) \in \mathscr{B} \right). \tag{EI}$$

In the EIV case, the fitting trajectory $\hat{w}$ is *any* trajectory of $\hat{\mathscr{B}}$, while in the EI case, the fitting trajectory $\hat{w}$ is $\text{col}(u, \hat{y})$, so that it is generated by the *given* input u and only the initial conditions are freely chosen.

Indeed, the EI case corresponds to output error identification, however, (EI) is maximum likelihood for *known* up to a scaling factor σ^2 output error covariance matrix $I\sigma^2$. If the output error covariance matrix is unknown, the maximum likelihood principle leads to the cost function

$$\det \left(\sum_{t=1}^{T} \big(y(t) - \hat{y}(t)\big) \big(y(t) - \hat{y}(t)\big)^\top \right),$$

which in the multi output case differs from the cost function of (EI).

Sections IV:A and B are extended and revised.

In IV:C make clear that the extra input ("latent variable") will play the role of innovations, and allowing a personal P-polynomial for this extra input essentially means that an ARMAX model is estimated. Using this e to make $w = \hat{w}$ is, as pointed out, the conventional prediction error method for ARMAX models.

The following clarification is added:

The unobserved input e, called *latent input*, plays the role of innovations. Written in a polynomial form, the model with latent inputs is the classical ARMAX model

$$P(\sigma)y = Q(\sigma)u + M(\sigma)e.$$

It is unavoidable that table I is read as a comparison of methods, even though you state on the page before that the last column of the table has to give the smallest M_{rel}, since that is what is minimized.

We revised the simulation results according the comments given below. Now the comparison is for the output error case. In addition, the data is split into 70% identification and 30% validation parts.

(1) Why is initialization time not included in the time for STLS, while it is included for PEM?

Initially the purpose of the simulation example was to show that the STLS method can be used as an "iterative refinement step" of subspace identification methods. For this purpose, we were interested in the time for the optimization step of the algorithm only. Later on we added results with the `pem` function but reused the available results for the STLS method. We agree that in this way the comparison is not fair. In the revised version of the paper, the initialization time is included in the computation times for both the STLS and prediction error methods.

(2) It is not clear (to us) which of the data sets are EIV and which are EI.

The data sets in DAISY are arrays of double precision numbers with a short description how they are obtained and what is the physical meaning of the variables. There is no clear indication whether the inputs are exact or perturbed. It is our subjective decision to make such an *assumption* and consequently apply EIV or output error identification. In the original version of the paper, all simulation examples were considered to be EIV identification problems, which certainly gave priority to the STLS method (it is designed for such problems) over the prediction error and N4SID methods (they are designed for ARMAX system identification). In order to make the comparison fare in this respect, all simulation results are redone as output error identification problems.

For the EI case, it would therefore be interesting to evaluate $\|y-\hat{y}\|/\|y\|$ also when `stls` has been computed to minimize $\|w-\hat{w}\|$. (Just like $\|w-\hat{w}\|/\|w\|$ is evaluated for `pem`, which minimizes $\|y-\hat{y}\|$.)

The STLS based method can minimize either $\|y-\hat{y}\|$ (EI setting) or $\|w-\hat{w}\|$ (EIV setting) depending on whether the input is specified as exact or not. Now the experiments are in the EI setting where both the STLS and PEM method are meaningful and the validation is performed in terms of the output error $\|y-\hat{y}\|/\|y\|$.

(3) Are `subid` and `detss` (page before table) for the EIV or EI case?

`subid` is designed for ARMAX system identification. `detss` is designed for exact system identification. Both they are applied heuristically to a situation where the assumptions under which they are derived are not satisfied. In the revised version, we excluded the exact identification method from the comparison but kept the N4SID method `subid` because it is used as an initial approximation for STLS.

(4) Recall in table caption what is T, m, p, l.

Done.

(5) The normal way to evaluate an identification result is to validate on validation data.

Done.

Reviewer #1:

1. The modeling is done in the behavior framework of Jan Willems, and I cannot judge its novelty

One of the main motivations for the formulation of the behavior framework is namely the need of a rigorous definition of what is meant by an exact and approximate model of an observed time series. The paper is certainly not original in this respect. Also the specific identification problem treated in the paper was proposed by Roorda and Heij in [RH95]. We are not original in this respect either. As the title of the paper states, our contribution is

> a link between work done in the field of numerical linear algebra and signal processing (the structured total least squares problem) and work on identification in the behavioral setting (the global total least squares problem).

The paper generalizes earlier results of De Moor [DM93] on identification of SISO systems to the MIMO case.

2. The method is evaluated on 24 DAISY datasets. The cost function presented is the error norm minimized by the proposed method. Hence it is clear that the STLS method gives the "best" results. No real model validation is performed.

We revised the simulation results. (See our response to the comments of the Editors.) Now the methods are validated on part of the data that does not overlap with the data used for identification.

There are no results on identification of nonlinear system!

The global total least squares method is designed for identification of LTI systems. However, as its primary objective is the *approximation* of the given data, it is well suited for identification of an LTI approximation of a nonlinear system. We do not pursue this aspect of the method in the paper but show a simulation example, see Section VI.C.

Reviewer #2

A minor error is: Page 10, line 2 of the Proof for Th. 3 : replace R by R^T.

Corrected.

1) On page 3, par. 4, lines 3–4, the authors claim that "Non-iterative methods ... solve the kernel problem via the singular value decomposition (SVD)". But the SVD algorithm is iterative, so techniques based on SVD are ultimately iterative. Maybe, the phrasing could be refined.

The reviewer is right. We meant to say that the solution is explicitly given in terms of the SVD factors. In addition, although the computation of the SVD is iterative, the classical SVD algorithms have global convergence to a global minimum point with cubic local convergence rate. The SVD is typically computed up to the machine precision in 5 to 10 iteration steps. Thus although in theory the SVD based methods are iterative, in practice they are very similar to non-iterative methods. We deleted "non-iterative" from the statement of page 3.

2) As defined on page 5 after formula (3), the σ operator, called "backwards shift operator", actually looks like a forward shift operator.

For $w \in (\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}}$,

$$w = \begin{bmatrix} \cdots & w(0) & w(1) & w(2) & w(3) & \cdots \end{bmatrix}$$
$$\sigma w = \begin{bmatrix} \cdots & w(0) & w(1) & w(2) & w(3) & \cdots \end{bmatrix}$$

The second line is a backwards shifted version of the first one, hence the name for σ—"backwards shift operator".

3) Page 5 , line -4 (bottom-up): it would be better to define the operator "col".

Done, in the place where the notation is first used (see Section II.C).

4) Page 10, line 2 before Th. 2: Theorem 2 is referred to as a "lemma".

Corrected.

5) Page 10, line 3 of Proof for Th. 3: it would be better to define the matrix $\mathscr{C}$.

Done.

6) Page 11, line 1 of subsection B: replace "systems" by "system".

Done.

7) Page 13, items 1-4: I think "Destillation" should be replaced by "Distillation".

Corrected.

8) Page 15, ref. 6: Some words seem to be missing in the title. Also, is "in In" OK (after the title)?

Corrected.

Reviewer #3

. . . I believe that no new result appears here which has not previously appeared elsewhere.

To the best of our knowledge, the application of the structured total least squares method for MIMO system identification, which is the main contribution of the paper, is original. Previous results on the application of the structured total least squares method for system identification [DM93, DR94, LD01] treat the SISO case only. In the conclusions of [DR94], the extension to the MIMO case is stated as an open problem.

the paper does not say anything about the identification of nonlinear systems

The global total least squares method is designed for identification of LTI systems. However, as its primary objective is the *approximation* of the given data, it is well suited for identification of an LTI approximations of nonlinear systems. We do not perused this aspect of the method in the paper but show a simulation example, see Section VI.C.

Last sentence of section I.D: Weighted norms can also be used in a non-stochastic setting to reflect prior knowledge. It may not be possible to give a precise interpretation of what is being done in this case, but it is common (and essential) practice.

The comment of the reviewer is added in the end of section I.D.

Just after conjecture 1: The word "motivation" is not the right one here. Perhaps "justification" is better.

Corrected.

Perhaps this conjecture is unnecessarily strong. Would it not be enough to show that every w is arbitrarily close to some $\hat{w}$ which is in $(R^{\mathfrak{w}})^T$? Essentially show that Ω is dense, rather than generic.

A dense set can be too "small". By generic set we mean a set that has an open, dense, and measure exhaustive subset. We conjecture that Ω is generic.

In section IV.C the authors define the phrase "the classical system identification framework" somewhat polemically, to suit their own ends. If order selection tests such as AIC, BIC, MDL, etc are considered to be included in the "classical framework" then the minimisation in (7) is not just over $\hat{e}$, but also over ℓ (max lag) and n_e (and not just of $\|\hat{e}\|$). In this case the "classical framework" differs from this paper in detail, rather than in philosophy.

We agree with the reviewer that the problem to determine the model complexity (the lag l and the number of latent inputs n_e) from the data is a problem in its own right and is not addressed in the paper. Our point of view is that l and n_e are user defined upper bounds on the model complexity. Whether or not certain choice of l and n_e allows good fit can be verified only after a model is identified.

One possible approach to the problem of determining l from the data (*i.e.*, misfit minimization only) is to plot the misfit of the optimal model as a function of l and choose from the plot a "good" value for l. The curve will typically have the "L" shape and a good value for the parameter is a one at the corner (small misfit achieved by a simple model). This is illustrated in the simulation example of Section VI.C.

Both misfit and latency are considered in the classical case, though the measures (and the language) used are different.

We will appreciate to know about specific references addressing the misfit minimization problem in the classical case.

I have problems with the phrase "the data is obtained from the errors-in-variables model" (section V.C). I can guess what the authors mean, but the notion of data arising from a model is highly confusing, particularly in a paper which includes fundamental issues in its scope.

The errors-in-variables model is

$$w = \bar{w} + \tilde{w}, \quad \text{where } \bar{w} \in \bar{\mathscr{B}} \text{ and } \tilde{w} \text{ is a stochastic process with known p.d.f.} \tag{EIV}$$

$\bar{\mathscr{B}}$ is the "true model", $\bar{w}$ is the true trajectory, and $\tilde{w}$ is the measurement noise. The phrase "the data w is obtained from the errors-in-variables model EIV" means that there is a true trajectory $\bar{w} \in \bar{B}$ and a realization $\tilde{w}$ of the measurement noise, such that $w = \bar{w} + \tilde{w}$. Our statement is similar to the often used statement "data generated by an ARMAX model".

The Conclusions imply that methods other than STLS cannot do certain things, such as model multivariate time series without a partitioning into inputs and outputs. This is clearly false, and the claims should be phrased more carefully.

We do not know which claims the reviewer refers to. The most relevant statement from the conclusion is

> "The STLS method allows to treat identification problems, without input/output partitioning of the variables ..."

which of course does not imply that the STLS method is *the only* one that can treat identification problems, without input/output partitioning of the variables. In fact, we cite in the paper other methods that solve this problem (*e.g.*, the global total least squares method of Roorda and Heij [RH95]). We do not really understand the objection of the reviewer.

References

[DM93] B. De Moor. Structured total least squares and L_2 approximation problems. *Lin. Alg. and Its Appl.*, 188–189:163–207, 1993.

[DR94] B. De Moor and B. Roorda. L_2-optimal linear system identification structured total least squares for SISO systems. In *Proc. of the Conf. on Decision and Control*, pages 2874–2879, 1994.

[LD01] P. Lemmerling and B. De Moor. Misfit versus latency. *Automatica*, 37:2057–2067, 2001.

[RH95] B. Roorda and C. Heij. Global total least squares modeling of multivariate time series. *IEEE Trans. on Aut. Control*, 40(1):50–63, 1995.