Author's response to the referees' reports on "A software package for system identification in the behavioral setting"

Ivan Markovsky

May 27, 2013

I thank the referees for their relevant and useful comments. In this document, I quote in **bold face** comments/questions from the reports. My replies follow in ordinary prints. In blue, I quote passages from the revised manuscript.

Reviewer #2

1. **It would be good to include**

   (a) **A comparison of the corresponding frequency responses for both estimated models and real system.**
   Done in all model reduction examples (Sections 5 of the revised version of the paper).

   (b) **Show the evolution of the parameter estimates with respect to the number of iterations.**
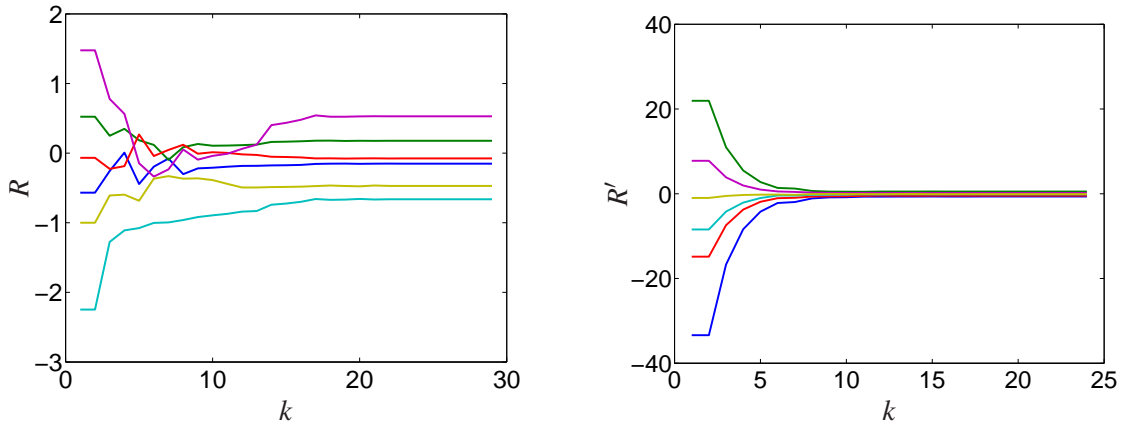   Done in Section 4 of the revised version of the paper.

Figure 1: Convergence of the model parameters R and R' in the course of the optimization method.

2. **In Section 6 the author addresses missing data problems. It would be good to analyze the results obtained in the simulation examples.**

   The following discussion of the results is added:

   The problems solved by the `ident` and `sysid` functions are equivalent. In the stochastic setting of Section 2.3, they are errors-in-variables identification problems with missing data. Examples 14 and 15 numerically confirm that the results computed by the two methods coincide for small noise levels and a

1

small number of missing values. Since the problems are nonconvex, however, for high noise levels the methods may converge to different locally optimal solutions.

In addition, there is an important algorithmic differences between `ident` and `sysid`, which explains the results of Example 16. In `ident`, the initial conditions are eliminated analytically at the level of the misfit computation (see Section 3). In `sysid`, the initial conditions are included in the parameter vector of the nonlinear least squares problem (see [PS00]). This results in a larger optimization problem, solved by `sysid` than the one solved by `ident`. Since in `sysid` a transient response is added for every missing data point, the number of initial conditions to-be-estimated is growing with the increase of number of missing data points. In contrast, the dimension of the nonlinear least squares problem solved by `ident` is fixed. The difference in performance between `ident` and `sysid` due to the different way of dealing with the initial conditions becomes pronounced in problems with more missing values. This is illustrated in Example 16, where the problem is no longer solvable by `sysid`. Detailed statistical and numerical analysis of the identification methods in the case of missing data will be presented elsewhere.

3. **The term "Multiple time series" has been extensively used in the Time Series Literature to represent multivariate time series. The author used this term with a different meaning. It would be good to change this term to e.g. "Time series from Multiple experiments".**

Thank you for pointing this out. The use of "multiple time series" is replaced by "multiple experiments".

4. **The Acronym "SYSID" should be defined where first appears.**

The use of "SYSID" as an acronym is now avoided. (Note that "SYSID" is still used as an equation label.)

Reviewer #3

• **The approximation criterion did not take into account the missing elements, it is a legibile approach but not justified or it must submit to some not cited conditions.**

Justification is given in Section 2.3 of the revised version of the paper, where the link between the identification problem considered and the errors-in-variables estimation problem is discussed. In the errors-in-variables setting, zero weights correspond to observations with infinite variance of the associate noise, which, informally corresponds to the notion of a missing value. Complete statistical analysis (consistency and efficiency of the estimator, error bounds, *etc*.) is outside the scope of the current paper and will be pursued elsewhere.

Reviewer #4

• **Absence of stochastic analysis / considerations: The whole problem is investigated only from the realization point of view and it is not shown how measurement or process noise in this setting play a role, especially in the case when IO partitions are not predefined.**

The relevant statistical setup for the deterministic approximation problem considered in the paper is the errors-in-variables setting. Section 2.3 explaining the link is added in the revised version of the paper.

The system identification problem considered in the paper is defined as a deterministic approximation problem. As shown in [Mar08, Section 3.1], however, it yields the maximum likelihood estimator in the errors-in-variables setting [Söd07].

Proposition 1 ([Mar08, Proposition 6])**.** *Assume that the data w_d is generated in the errors-in-variables setting*

$$w_d = \bar{w} + \widetilde{w},$$

2

where the true data $\bar{w}$ is a trajectory of a true model $\bar{\mathscr{B}} \in \mathcal{L}_{\mathtt{m},\ell}^{\mathtt{w}}$ and the measurement noise $\widetilde{w}$ is zero mean normally distributed with covariance matrix that is a multiple of the identity. Then the solution $\widehat{\mathscr{B}}^*$ of (SYSID) is a maximum likelihood estimator for the true model $\bar{\mathscr{B}}$.

In [PS01, KMV05], it is proven that under additional mild assumptions the estimator $\widehat{\mathscr{B}}^*$ is consistent and the estimated parameters have asymptotically normal joint distribution. Therefore, asymptotic confidence regions can be obtained as a byproduct of the optimization algorithm (see [MU12, Section 5.1]). The statistical setting justifies the choice of the deterministic approximation criterion and provides a testbed for the method: the method works "well" (consistency) and is optimal (asymptotic efficiency) under specified conditions.

Note 2 (Exact and missing data correspond to noisy data with, respectively, zero and infinite noise variance). In the classical regression model, an input-output partitioning of the variables is a priori given and the input variables (regressors) are assumed to be noise free. Classical regression is a special case of the errors-in-variables model when the input noise variance is zero. The opposite of a noise free variable is a missing variable, or equivalently a variable with infinite variance in the errors-in-variables setting.

- **The considered system is quasi linear due to the constant terms. This does not fit to the considered system class in terms of (DE).**

Constant terms in the differential equations (DE) result in an offset of the output. They can be incorporated in the classical linear time-invariant setting by augmenting the system with a pole at 1. Let $\mathscr{B}$ be a linear time-invariant system of order $\mathtt{n}$ without a constant term in the output. Then the system

$$\mathscr{B}' = \left\{ \begin{bmatrix} u \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ y_{\text{const}} \end{bmatrix} \ \middle| \ \begin{bmatrix} u \\ y \end{bmatrix} \in \mathscr{B} \right\}, \tag{1}$$

where

$$y_{\text{const}} = (c_{\text{offset}}, c_{\text{offset}}, \cdots)$$

is a linear time-invariant system of order $\mathtt{n}+1$. Indeed, if

$$\mathscr{B} = \mathscr{B}(A, B, C, D), \qquad x(0) = x_{\text{ini}}$$

is a state space representation of $\mathscr{B}$, then $\mathscr{B}'$ admits a state space representation

$$\mathscr{B}' = \mathscr{B}\left(\begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} B \\ 0 \end{bmatrix}, \begin{bmatrix} C & c_{\text{offset}} \end{bmatrix}, D \right), \qquad x'(0) = \begin{bmatrix} x_{\text{ini}} \\ 1 \end{bmatrix}. \tag{2}$$

- **Secondly, why do these terms appear at all?**

In the original version of the paper, I forgot to give the definition of the mass position p_i. This is done in the revised version:

Denoting by p_i the position of the ith point mass with respect to the left wall, ...

The definition of the p_i's with respect to a *common reference point* is a reason for the appearance of the constant terms in (DE). The suggestion of redefining the p_i's with respect to the corresponding equilibrium positions is discussed in the answer to the next comment.

Independent of the definition of the positions p_i, however, output offsets (constant terms) appear when the chain of masses is not connected by a spring to the walls (see Example 8 in the paper.) In this case, the reason is an intrinsic feature of the system's behavior rather than an artificial one caused by "unsuitable" choice of the model representation and can not be eliminated without changing the problem.

- **If you chose all positions as local deviations from given reference points described by the nominal length of the springs, then these constants and the related virtual state variables are superfluous.**

  If the system has an equilibrium position (this is the case when the chain of masses is connected to a wall by a spring), it is indeed possible to avoid the state $x_{n+1} = 1$ in (2). By doing this, however, one makes the spring length (a parameter of the model) "externally" specified. It will no longer be a specification of the signal behavior $\mathscr{B}$. In an identification context, this implies that a physical parameter is a priori known or it is independently estimated. Both options have drawbacks. The former leads to a loss of generality. The latter splits the identification problem into two independent steps: estimation of the spring length and estimation of the remaining model parameters. Therefore, the solution is suboptimal compared with the simultaneous estimation of all parameters. Dealing with offsets in the outputs is a more general approach that avoids these shortcomings.

  In the case when the chain of masses is detached from the walls (Example 8 in the paper), the model has intrinsic offset due to a free motion of the whole chain. Then, the offsets can not be avoided by redefining the positions or centering the data and one has to identify a behavior of the type (1).

- **Page 1, Item 1: In practice, experiments are usually conducted by applying excitation at specific "ports" of the system. Therefore, it is doubtful if leaving the a priori information regarding this cause and effect relation out of the identification problem will be beneficial or not. This undermines the motivation presented under Item 1.**

  If the approximation criterion is symmetric in the variables (errors-in-variables setup), the optimal approximation error is not affected by a permutation of the variables. (A numerical confirmation is included in Section 4.) This shows that a priori information of the input/output partition of the system used to generate the data is irrelevant. In the output error setting, the partitioning of the variables is important because the criterion is asymmetric. Using asymmetric criterion is useful when there is prior knowledge that some variables (*e.g.*, the inputs) are exact. This does not undermine the point of item 1 in the paper. It just articulates the importance of taking into account prior knowledge about the data. Prior knowledge about exact variables can be used independent of the model representation and need not be linked to an input/output representation.

- **page 2, line 3: vector observation -> observation vector**

  Corrected.

- **Page 3: "The behavioral approach and the low-rank approximation setting fit like a hand fits a glove." This sentence is a bit too colorful and shallow at the same time.**

  The paragraph in which the sentence appears is rephrased as follows:

  The behavioral approach and the low-rank approximation setting are two sides of the same coin: they use representation free problem formulations, where the fundamental object of interest is a relation among variables rather than a map. The system theoretic notion of a linear time-invariant model of bounded complexity corresponds to the linear algebra notion of a rank deficient Hankel matrix. The linear algebra setting provides algorithms and software for addressing problems in the system theoretic setting.

- **Page 4: In the block partitioning of R_l please reveal the connection of Q_l and P_l with the chosen IO partition.**

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

  No a priori separation of the variables $w_1, \ldots, w_q$ into inputs and output is made. However, the variables can always be partitioned into inputs u (free variables) and outputs y (dependent variables). A convenient way to represent an input/output partitioning is $w = \Pi \begin{bmatrix} u \\ y \end{bmatrix}$, where Π is a $q \times q$ permutation matrix.

  An input/output representation

  $$\mathscr{B}(P, Q, \Pi) = \{ w = \Pi(u, y) \mid Q(\sigma)u = P(\sigma)y \},$$

of the system $\mathscr{B}(R)$ is obtained by partitioning the polynomial matrix $R\Pi$ as

$$R\Pi = \begin{matrix} \mathtt{m} & \mathtt{p} \\ [Q & -P] \end{matrix}, \quad \text{with } P \text{ nonsingular.}$$

In addition, the system can be represented in the input/state/output form

$$\mathscr{B} = \mathscr{B}(A,B,C,D,\Pi) := \{\, w = \Pi(u,y) \mid \text{there is } x, \text{ such that } \sigma x = Ax + Bu,\ y = Cx + Du \,\}. \quad \text{(I/S/O)}$$

If the permutation matrix Π is the identity matrix, then $w = (u,y)$, *i.e.*, the first $\mathtt{m}$ variables of w are inputs and the remaining $\mathtt{p}$ variables are outputs. In this case, Π is skipped in (I/S/O), *i.e.*, $\mathscr{B}(A,B,C,D) = \mathscr{B}(A,B,C,D,I)$.

A note is added to make a link with the term "kernel representation" in the literature on the behavioral approach.

Note 3 (Kernel representation). In the literature on the behavioral approach to systems and control, (DE) is called a kernel representation, because it can be written more compactly as the kernel

$$\ker\big(R(\sigma)\big) := \{\, w \mid R(\sigma)w = 0 \,\}$$

of the operator $R(\sigma)$, where

$$R(z) := R_0 w + R_1 z + \cdots + R_\ell z^\ell$$

is a polynomial matrix.

- **Equation (M): This ID criterion also allows to treat error-in-the-variable problems.**

  Yes, the link to error-in-variables identification is explained in Section 2.3 of the revised version of the paper.

- **Page 6: $\mathscr{B}_{\mathrm{aut}}$ above Proposition 3 is undefined. Would not be better to call it as the autonomous part of behavior $\mathscr{B}$?**

  Yes, thank you for the suggestion: "related autonomous system $\mathscr{B}_{\mathrm{aut}}$" has been replaced by "the autonomous part of behavior $\mathscr{B}$".

- **Page 10: Section 7: The Author first mentions that no tuning is applied on the hyper parameters as an identification process should be as automatic as possible. Just in the following paragraph this is continued by describing how the tuning with respect to ℓ was done. Isn't this a contradiction?**

  The specification of ℓ is a part of the problem formulation (see (SYSID)) and is explicitly mentioned:

  > The methods are applied on all data sets choosing only the model class (specified by a bound on the model complexity).

- **Wouldn't be better to see it as a model structure selection step which must be also done in the PEM case (implemented in the procedure itself)? To see better the number of estimated parameters please also give the model structures obtained by PEM for the listed data sets.**

  Yes, the ℓ parameter defines the model structure and can be chosen by model order selection methods. The values determined automatically by `pem` are as follows:

example #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
ℓ manual	1	1	2	1	5	2	5	4	1	2	6	1	2	2	1	1	1
ℓ pem	–	–	1	2	1	3	2	6	4	1	2	1	1	3	1	2	1

  In the first two data sets the number `pem` issues an error message:

```
There are too many parameters to estimate for chosen estimation data size.
```

References

[KMV05] A. Kukush, I. Markovsky, and S. Van Huffel. Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. *J. Statist. Plann. Inference*, 133(2):315–358, 2005.

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

[MU12] 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.

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

[PS01] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach*. IEEE Press, Piscataway, NJ, 2001.

[Söd07] T. Söderström. Errors-in-variables methods in system identification. *Automatica*, 43:939–958, 2007.