System identification with missing data via nuclear norm regularization

Cristian Grossmann, Colin N. Jones, Manfred Morari

June 4, 2009

Abstract— **The application of nuclear norm regularization to system identification was recently shown to be a useful method for identifying low order linear models. In this paper, we consider nuclear norm regularization for identification of LTI systems from data sets with missing entries under a total squared error constraint. The missing data problem is of ongoing interest because the need to analyze incomplete data sets arises frequently in diverse fields such as chemistry, psychometrics and satellite imaging. By casting the system identification as a convex optimization problem, nuclear norm regularization can be applied to identify the system in one step, i.e., without imputation of the missing data. Our exploratory work makes use of experimental data sets taken from an open system identification database, DaISy, to compare the proposed method named NucID to the standard techniques N4SID, prediction error minimization and expectation conditional maximization via linear regression. NucID is found to consistently identify systems with missing data within the imposed error tolerance, a task at which the standard methods sometimes fail, and to be particularly effective when the data is missing with patterns, e.g., on multi-rate systems, where it clearly outperforms existing procedures.**

I. INTRODUCTION

The need to identify a dynamic system from an incomplete data set is a rather common situation in practice. There are different reasons that lead to missing entries in the data sets available for identification, such as: Sensor failures, outliers or plant shutdowns, which generate missing entries in the data set at random and multi-rate sampling or periodic disturbances that create patterns of missing data. Over the last three decades a number of researchers from various fields have recognized the need for systematic methods to exploit incomplete data sets for system identification and it is still recognized as a big and open challenge in process industry [1].

The goal of this paper is to present a recently developed method for system identification from noise-corrupted data with missing entries in the outputs. The proposed method is applicable to SISO and MIMO systems, identifies a non-parametric linear model and incorporates the minimization of the order of the identified system in a natural and transparent way by approximating it with the nuclear norm, i.e., by the sum of the singular values, of the Hankel matrix built from finite impulse response (FIR) coefficients. The resulting nuclear norm regularization for the rank of a matrix is the analogue to the l_1 regularization for vector cardinality, which is a well-known heuristic that produces sparse solutions. These regularization methods have been studied in detail by a number of researchers and set the foundation of the recently developed compressed sensing frameworks for measurement, coding and signal estimation [2], [3], [4].

The proposed technique minimizes the nuclear norm of the Hankel matrix of FIR coefficients while constraining the fitting error between model and data to a desired level of accuracy. This method allows one to directly choose a desired accuracy and then poses a convex optimization problem to find the lowest order model that achieves it, rather than iteratively tuning the order of the model, as is common practice. Nuclear norm regularization has been recently suggested by [2], [6] as a way to promote the identification of low order models out of *complete data sets*. This work shows how the nuclear norm regularization is specially attractive, when the data sets have missing entries, i.e. for *the missing data problem.*

A sensitivity analysis of the identification algorithm is performed on different structures of missing data in the outputs: structured missing data and randomly distributed missing data. The proposed method is compared under these scenarios to commonly used methods for identification with missing data and several case studies are performed on experimental data sets taken from the DaISy Database [9].

NucID is found to consistently identify systems from complete data sets or data missing at random within the imposed error tolerance, a task at which the standard methods sometimes fail. In the case of structured missing data, NucID is shown to be particularly effective and clearly outperform existing procedures. This poses NucID as an attractive tool for the identification of multi-rate sampled-data systems.

The paper is organized as follows: The general identification problem and the identification problem with missing data are defined in Section II and IV, respectively. Section III describes the nuclear norm regularization. The methods for comparison and the results of the identification of experimental data sets are presented in Sections V and VI.

II. PROBLEM FORMULATION

The identification problem is first formulated for the case where no data is missing in the outputs, before being extended in section IV to the general case of missing data.

The goal is to identify a discrete-time linear time-invariant model of the lowest possible order that can explain a sequence of input $u(t) \in \mathbb{R}^m$ and output measurements $y^{meas}(t) \in \mathbb{R}^p$ over an observation window $t = 0, \ldots, N-1$. We use the shorthand matrix notation for inputs $U \in \mathbb{R}^{N\times m}$ and outputs $Y^{meas} \in \mathbb{R}^{N\times p}$ by stacking the vectors $y^{meas}(t)$ and $u(t)$ rowwise. No assumptions on the specific structure or order of the model are made and the output i at time instance t, i.e., $y_i(t)$, is represented as a linear combination of the impulse responses of the inputs $j = 1, \ldots, m$, i.e., through a finite impulse response (FIR) model

$$y_i(t) = \sum_{j=1}^{m} \sum_{\tau=t-r}^{t} h_{ij}(t-\tau)u_j(\tau) + v_i(t) \quad i = 1, \ldots, p \tag{1}$$

The values h_{ij} are the FIR coefficients from input j to output i and the zero-mean white-noise $v_i(t)$ captures the unmeasurable disturbance affecting output i at time t. The sequence of FIR coefficients for channel i, j has length r,

All autors are at the Automatic Control Laboratory of ETH Zurich, 8092 Zurich, Switzerland.

{grossmann,cjones,morari}@control.ee.ethz.ch

which is a parameter that must be chosen large enough to describe the dynamics of the system to be identified.

The total squared error in the identification procedure e_N can be quantified by the sum of the squared differences between the measurements Y^{meas} and the outputs Y predicted by model (1) over the N samples:

$$e_N := \sum_{t=0}^{N} (y^{meas}(t) - y(t))^2 = \|Y^{meas} - Y\|_F^2 \ , \qquad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm.

The FIR coefficients $h_{ij}(t)$ for $t = 0, \dots, r$ of each of the $i \cdot j$ channels of model (1) are the variables to be estimated in order to describe the set of data Y^{meas} within a given error bound $e_N \leq \gamma$. The order of the resulting model is given by the rank of the Hankel matrix $\mathcal{H}_h$ formed from the impulse response coefficients h_{ij}

$$\mathcal{H}_h := \begin{bmatrix} h(0) & h(1) & \cdots & h(r-n_H) \\ h(1) & h(2) & \cdots & h(r-n_H+1) \\ h(2) & h(3) & \cdots & h(r-n_H+2) \\ \vdots & \vdots & & \vdots \\ h(n_H) & h(n_H+1) & \cdots & h(r) \end{bmatrix} \qquad (3)$$

where each entry $h(t)$ is a matrix in $\mathbb{R}^{p \times m}$ containing the coefficients $h_{ij}(t)$ of all channels for the corresponding time step t, $n_H := r/2$ and r is assumed to be even. Note that as long as r is long enough compared to the system dynamics, the order of the identified model has nothing to do with r. The order of model (1) can be understood as the number of states of the corresponding state-space model.

The search for a model of the lowest order that satisfies the error bound $e_N \leq \gamma$ can be posed as the following optimization problem:

$$\begin{aligned} \min_h \quad & \operatorname{rank}(\mathcal{H}_h) \\ \text{s.t.} \quad & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned} \qquad (4)$$

Alternatively, problem (4) can be written as

$$\min_h \|Y^{meas} - Y\|_F^2 + \alpha \operatorname{rank}(\mathcal{H}_h) \qquad (5)$$

in which the trade-off between the quality of fit and the order of the model is made explicit i.e., a Pareto curve can be obtained by varying α.

III. MINIMUM-RANK MODELS VIA NUCLEAR NORM MINIMIZATION

Minimizing the rank of a matrix $A \in \mathbb{R}^{n \times n}$ is a nonconvex problem and is in general NP-hard. The *nuclear norm* is a convex heuristic for rank minimization that was proposed in [5] and shown in [2] to be the *convex envelope*, or the closest convex function to the rank operation:

$$\|A\|_* := \sum_{i=1}^{n} \sigma_i(A) \qquad (6)$$

where $\sigma_i(A)$ is the i^{th} singular value of A. In the last few years, minimization of the l_1 norm has been used as a convex approximation of cardinality minimization, or to promote sparsity in the decision vector of optimization problems, in fields ranging from statistics [7] to communications [4]. Since the singular values of a matrix are all positive, the nuclear norm of A is equal to the l_1 norm of the vector formed from the singular values of A. As a result, minimizing the nuclear norm (6) will lead to sparsity in the vector of singular values, or equivalently to a low-rank matrix A.

We now turn to the optimization problem (4) and relax the non-convex rank to a nuclear norm minimization:

$$\begin{aligned} \min_h \quad & \|\mathcal{H}_h\|_* \\ \text{s.t.} \quad & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned} \qquad (7)$$

The above optimization problem can be re-cast as a semi-definite program (SDP) [5]

$$\begin{aligned} \min \quad & \mathbf{tr}\,(V_1) + \mathbf{tr}\,(V_2) \\ \text{s.t.} \quad & \begin{bmatrix} V_1 & \mathcal{H}_h^T \\ \mathcal{H}_h & V_2 \end{bmatrix} \succeq 0 \\ & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned} \qquad (8)$$

where we introduce the symmetric matrices V_1, $V_2 \in \mathbb{R}^{n_H \cdot p \times n_H \cdot p}$ as decision variables. Optimization problem (8) can therefore be posed and solved using standard SDP software (e.g., [8]).

Computational complexity: The SDP (8) has a large number of variables due to the introduction of the matrices V_1 and V_2, which limits the scale of problems that can be solved. In [6] a custom interior point solver for a related class of SDPs was proposed that offers speed improvements of orders of magnitude over previous algorithms and should be applicable to the SDP (8) with minor modification. The method [6] was used for system identification without missing data, but the technique is based on minimizing the nuclear norm of $Y^{meas}U^{\perp}$, which requires a significantly larger number of optimization variables than the proposed cost $\|\mathcal{H}_h\|_*$.

IV. SYSTEM ID WITH MISSING DATA

A. Problem formulation with missing data

We assume that all inputs have been sampled at a constant rate and that they are all available, i.e., we have N inputs $u(t)$ for $t = 0, \dots, N-1$ that, as before, can be collected in a matrix $U \in \mathbb{R}^{N \times m}$. Given the FIR model h, we can then write a linear function of h and U (1) to compute the matrix $Y \in \mathbb{R}^{N \times p}$, which is the *predicted* output of the model at all sample points $t = 0, \dots, N-1$.

In the case of missing data not all samples $y_i^{meas}(t)$ will be measured. The available outputs are recorded rowwise in a *measurement* output matrix $Y^{meas} \in \mathbb{R}^{\tilde{N} \times p}$. Note that Y^{meas} contains fewer entries than Y, i.e., $\tilde{N} < N$, because only the points in time with available measurements of the predictions Y are stored in Y^{meas}. In order to make these two matrices comparable, we define a measurement matrix $M \in \mathbb{R}^{\tilde{N} \times N}$ that maps the predictions onto the space of available measurements, $M : \mathbb{R}^{N \times p} \mapsto \mathbb{R}^{\tilde{N} \times p}$. In the case where all measurements are available, M is simply the identity matrix I.

As before, the error e_{MD} under missing data is defined as the sum of the squared differences between the predictions MY at the points in time where data is available, and the measurements Y^{meas}

$$e_{MD} := \|Y^{meas} - MY\|_F^2 \ . \qquad (9)$$

Standard approaches for fitting models with missing data first generate the missing measurements by interpolating the available data Y^{meas} and then use regular model identification techniques. The limitation of these approaches is that they must make an assumption on how this data is to be interpolated. Here, we make no such assumptions and consider fitting the data only at the measured points. The minimization of the nuclear norm can then be thought of as an interpolation method for the missing data where the interpolation is done by fitting a function in the class of low-rank dynamic systems.

Identifying a low-order model of the form (1) within a given error bound γ_{MD} from the incomplete data set U and Y^{meas} can now be cast as the convex optimization problem

$$\begin{aligned} \min_{h} \quad & \|\mathcal{H}_h\|_* \\ \text{s.t.} \quad & \|Y^{meas} - MY\|_F^2 \leq \gamma_{MD} \end{aligned} \tag{10}$$

A sensitivity analysis was carried out on problem (10) to investigate the effect on the identified dynamical model of different measurement matrices M, i.e., different patterns and amounts of output missing data. Two cases were investigated: (a) The missing output entries repeat themselves with the same pattern along the output matrix Y^{meas} and, (b) The missing output entries are randomly distributed along the output matrix Y^{meas}. In both cases we assume that all inputs are available.

Remark 1: The measurement matrix M as defined above assumes that all the outputs $j = 1, \ldots, p$ will be missing at the same time instance t . This rather restrictive and unrealistic assumption can be dropped in a straightforward way by defining a measurement matrix M_j for each output channel $j = 1, \ldots, p$. The formulation remains as in problem (4) by redefining the errors as

$$e_{MD} := \sum_{j=0}^{p} \|Y_j^{meas} - M_j Y\|_F^2 \tag{11}$$

B. Structured missing data

Sensors and actuators can have different rates at which they acquire data or take setpoints, respectively. In this work we consider the case where sensors and actuators work synchronously but at different rates. This can be interpreted as a multi-rate process between inputs and outputs, or amongst different outputs. This case corresponds to building the measurement matrix M by retaining only every n^{th} row of an identity matrix. Note that multi-rate scenarios lead very quickly to high percentages of missing data $MD_{\%}$, e.g., the simplest case where every second measurement of the outputs is not recorded corresponds to a percentage of missing data of $MD_{\%} = 50\%$.

C. Randomly missing data

Problems in sensors during acquisition can lead to loss in the measured data at random points in time. Different percentages of missing data $MD_{\%}$ have been considered, ranging from no missing data, $MD_{\%} = 0\%$ to $MD_{\%} = 70\%$. The measurement matrix M in this case is built by randomly dropping rows from an identity matrix with a uniform distribution.

V. NUMERICAL EXAMPLES

The proposed identification method, from now on referred to as nuclear norm identification (NucID), was compared with standard toolboxes available in MATLAB. Real experimental data from the DaISy Database [9] was used for the following identification experiments:

1) No missing data. The complete data sets were used to identify a linear dynamic model.
2) Structured missing data. The outputs are sampled at a lower rate than the inputs.
3) Random missing data. Some percentage $MD_{\%}$ from the output measurements is lost at random.

The outputs from the data sets were removed according to the pattern chosen for each experiment and are specified in each case.

A. Benchmark methods

Three different identification techniques were chosen for comparison with NucID. The corresponding MATLAB toolbox is given in brackets.

1) N4SID: Estimate a state-space model using subspace identification techniques. (n4sid)
2) PEM: Estimate a state-space model using an iterative prediction-error minimization method. (pem)
3) Expectation Conditional Maximization using Linear Regression (LR): Estimate a FIR model using multivariate linear regression with missing data. (ecmmvnrmle)

At this point it is important to note the way these methods are used when data is missing. In principle, there are two options: The missing entries can be simply disregarded in the identification procedure or one can try to guess the values of the missing entries, which is known as imputation. There are different techniques to *impute* the values of the missing data, e.g., linear interpolation, regression imputation, expectation maximization.

MATLAB offers the toolbox ‘misdata’ to impute the value of missing entries of data sets. The algorithm alternates between estimating models with N4SID from the available data and estimating missing data points. This iterative procedure is repeated until a given relative tolerance is achieved (1%) or for a maximum number of times (10 by default). The “reconstructed” data set can then be used with the three identification methods N4SID, PEM and LR.

This two step procedure of imputing values of missing entries and then identifying a model does not apply for the NucID method, which is a one step procedure that does not need any imputation of the missing values. This is one of the key benefits of the proposed method, since the procedure of imputing the data will often either cause a significant artificial increase in model order, or will generate nonsensical results when large percentages of data are missing.

VI. RESULTS

This section presents the identification of several dynamical systems comparing the proposed method with three standard identification tools, N4SID, PEM and LR, presented in the previous section. First the identification problem of a CD player arm is presented and discussed in detail to highlight the advantages of NucID over N4SID, PEM and LR. Then the same detailed analysis for a number identification problems is summarized in section VI-B.

A. Identification of a CD player arm

The experimental data from a mechanical construction of a CD player arm is considered here. The system has two inputs that are forces of the mechanical actuators and two outputs that are related to the tracking accuracy of the arm. The data set contains $2,000$ sample points out of which 400 were used for the identification procedure and the rest to validate the identified models. In the first experiment all measurements were considered while for the second and the third experiments data was dropped from the outputs according to the strategy described. The error presented in this paper has been normalized with the factor $|Y_I - \bar{Y}_I|_2$, where Y_I is the complete set for identification and $\bar{Y}_I$ the mean of the samples.

1) Complete data set: The complete data set was used to identify a dynamical model using N4SID, PEM, LR and NucIDand the resulting impulse responses are presented in Fig. 1. For the sake of clarity, the impulse responses for LR were not plotted since they are of high order and make the figure unclear, as illustrated further on. The models identified with N4SID and PEM have an order of two and are in general in good agreement with the FIR coefficients identified by NucID.

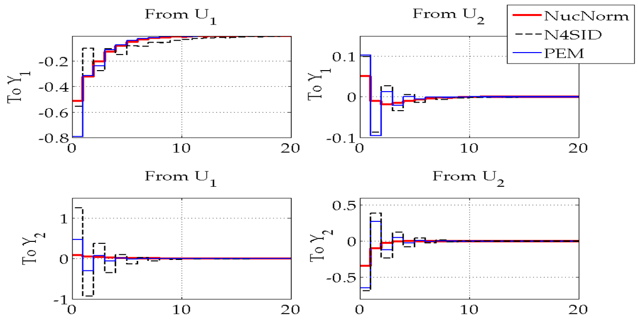

Fig. 1. FIR coefficients of the identified dynamical system for NucID, N4SID, PEM using the complete data set.

In order to provide an overview of the order of the model identified by NucID, a singular value decomposition (SVD) of the Hankel matrix built from the FIR coefficients $\mathcal{H}_h$ was computed and inspected. The order was then defined as the number of singular values above 0.01% (10^{-4}) of the first value. Fig. 2 shows the SVD of $\mathcal{H}_h$ for five NucID identification procedures using different values for the error bound γ. It can be observed that by decreasing γ, the number of non-negligible singular values (denoted with a star) increases, i.e., the order of the identified model is higher.

The four approaches are compared by plotting for each method the order of the identified models against the corresponding normalized error in Fig. 3. For the methods N4SID and PEM, models with fixed orders from 1 to 10 were identified and their normalized errors computed. The LR method yields only one point, since there is no way to choose the order of the identified model as in the other methods. For the NucID method the tuning of the order is done through varying γ and the order and errors of five runs presented in Fig. 2 for different values of γ are plotted in Fig. 3.

It is evident for N4SID, PEM and NucID, that there is a trade-off between the order of the identified model and the prediction error. The NucID method is able to identify models that give lower prediction errors than models identified with N4SID and PEM with the same order. It is well known that LR gives a rather good fit, but with very high order models, as illustrated in Fig. 3.

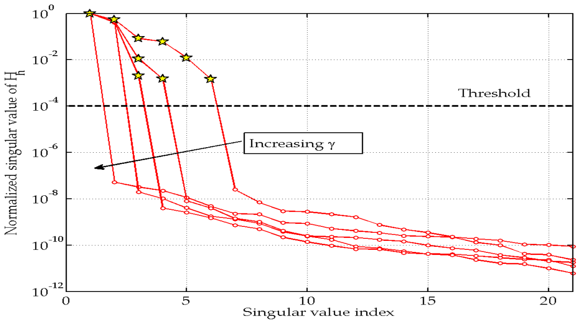

Fig. 2. SVD of the Hankel matrix $\mathcal{H}_h$ built from the FIR coefficients identified with the NucID method for different error bounds $\gamma = 2.32,\ 2.35,\ 2.45,\ 2.75,\ 7.8$. The stars define the non-negligible singular values above the threshold of 10^{-4}, that define the order of the model.

We can conclude that when using the complete set of data the NucID method is able to identify dynamical models that are comparable to the other methods in terms of model order and prediction error, and that in this specific case slightly outperforms N4SID, PEM and LR. Similar results for the identification of low-order models from complete data sets had also been observed by [6]. The next step is to assess the impact of missing output data on the identified models with the different methods.

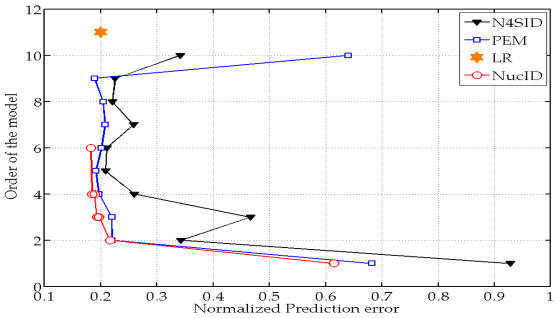

Fig. 3. Order of the identified model as a function of the normalized prediction error for NucID, N4SID, PEM and LR.

2) Structured missing data: This section presents the identification of the CD player arm example assuming that the output data was collected at a slower sampling rate than the one of the inputs, as encountered in a multi-rate sampled-data process. In the first example, we consider four experiments, runs (a) – (d), where the measurements of *only* the first output is sampled at a lower frequency than the inputs, and therefore the percentage of missing data in the outputs does not exceed 50%. Note that for the first example, the second output of runs (a) – (d) is sampled at the the same rate as both inputs, i.e., the data set still contains all the output measurements of the second output. In a second example, runs (e) – (h), both channels are sampled at lower rates, which results in higher percentages of missing data. The sampling rates that were analyzed are reported in Table

TABLE I

Multi-rate identification experiments of the CD player arm

Run	Normalized sampling time of				$MD_{\%}$
	Input 1	Input 2	Output 1	Output 2	
(*)	1	1	1	1	0 %
(a)	1	1	2	1	25.00%
(b)	1	1	3	1	33.33%
(c)	1	1	4	1	37.50%
(d)	1	1	8	1	43.75 %
(e)	1	1	2	8	68.75 %
(f)	1	1	8	2	68.75 %
(g)	1	1	3	8	77.08 %
(h)	1	1	8	8	87.50 %

I. The last column of Table I is the percentage of missing data in the outputs $MD_{\%}$ compared to run (*), that corresponds to the case analyzed in the previous section, where the entire data set is available for identification. The results for runs (a) – (d) are plotted in Fig. 4. The NucID and the LR method identify models with rather small prediction errors, around 0.2 for runs (a) – (c). These values are comparable with the ones from run (*) in Fig. 3. For run (d) the performance of the LR method deteriorates considerably while the model identified with the NucID method is the same as before. The order of the identified models by LR are substantially higher (11th order) than the ones from NucID (4rd order).

Models of different orders identified with N4SID and PEM in each run are connected with a line in Fig. 4. We can observe how their performance greatly deteriorates as the amount of missing data increases from $MD_{\%} = 25\%$ (run (a), solid line —) to $MD_{\%} = 33.3\%$ (run (b), dotted line $\cdots$) and $MD_{\%} = 43.75\%$ (run (d), dashed line - - -). For the sake of clarity, run (c) has not been shown, but it follows the same trend.

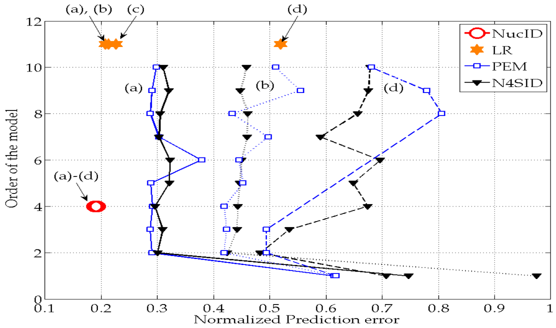

Fig. 4. Runs (a)-(d): Order of the identified model as a function of the normalized prediction error for NucID, N4SID, PEM and LR for structured missing data. Models of different orders identified with N4SID and PEM in the same run are connected with a line; Run (a): solid (—). Run (b): dotted ($\cdots$). Run (d): dashed (- - -). For the sake of clarity, run (c) has not been shown.

In a second example, we consider the case were both outputs are sampled at lower rates than the inputs. Four different scenarios are studied. Runs (e) and (f) have the same amount of missing data, but the sampling rates are exchanged between the outputs to test the sensitivity of the identified model to the output with more or less missing data. In runs (e) and (f) the measurements of the outputs are collected synchronously, but the sampling rates differ by a factor of four. Runs (g) and (h) investigate rather high percentages of output missing data.

The results for the second example are presented in Fig. 5. Note that the scale on the x-axis has changed from Fig. 4 to Fig. 5. Only LR and NucID were able to find a solution in all runs. LR identified high order models (11th) with larger errors than in the previous examples. In runs (e) and (f), N4SID and PEM identified models that predicted unacceptably large errors. These methods are also sensitive to the output with higher or lower sampling rates, since the models identified in run (e) (solid line —) have much smaller errors than in run (f) (dashed line - - -). For runs (g) and (h), N4SID and PEM were not able to find stable models. The NucID method consistently identified in runs (e) –(h) the same model as in all the previous runs, with the same order (4th order) and prediction error (approx. 0.2). A good consistency of NucID is was observed throughout the runs (a) – (h) where the identified FIR coefficients were virtually identical to those identified from the complete set of data in run (*).

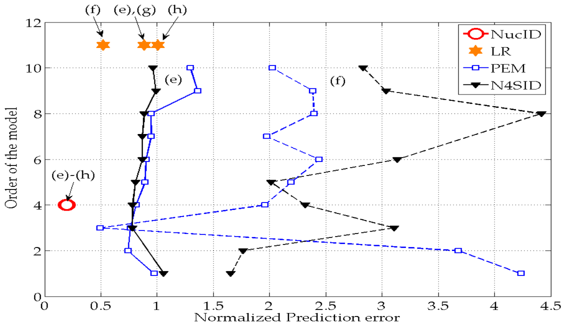

Fig. 5. Runs (e) – (h): Order of the identified model as a function of the normalized prediction error for NucID, N4SID, PEM and LR for structured missing data. Models of different orders identified with N4SID and PEM in the same run are connected with a line; Run (e): solid (—). Run (f): dashed (- - -). N4SID and PEM did not identify a stable model for runs (g) and (h).

3) Randomly missing data: In this example, an increasing percentage of the output entries $MD_{\%}$ is missing at random throughout the measurements and the results are reported in Table II, where each row represents a different amount of missing data. The normalized errors e_I for each of the methods is reported together with the order n of the identified model. For NucID, N4SID and PEM the identification error of the same order models are reported, whereas for LR the errors correspond to higher order models. After 50% of missing data N4SID and PEM fail to identify a model, indicated by a star *. Only NucID and LR are able to identify a model when more than 40% of the data is missing, although only NucID finds a model of a reasonable order.

NucID, PEM and LR give normalized errors in the same range for up to 45% of missing data, although of course the order of the LR models is much higher. The performance of N4SID is acceptable only for some specific instances and after 50% of missing data, N4SID and PEM fail to identify a model at all. The NucID method is able to identify the system with up to 75% of missing data with rather small errors.

B. System identification from DaISy database

In this final section we present selected scenarios from the previous analysis applied to different systems taken from the

TABLE II

RESULTS FOR MISSING DATA AT RANDOM

		NucID	N4SID	PEM	LR	
$MD_{\%}$	n	e_I			n	e_I
10	3	0.1965	0.2490	0.2062	11	0.2042
20	3	0.1995	0.5682	0.2054	11	0.2066
30	3	0.1970	0.7576	0.2119	11	0.2080
40	5	0.1912	0.2265	0.2025	11	0.2134
50	4	0.1934	*	*	11	0.2571
60	4	0.1935	*	*	11	0.2900
70	3	0.2303	*	*	11	0.3867

database for system identification (DaISy) [9]. Due to space restrictions, we present only the main results in Table III. The type of system, reference number and the number of inputs and outputs (Inputs x Outputs) can be found in the first column. In all the results of this section, two scenarios for each system are presented, one with the complete data set and one multi-rate scenario as presented in section VI-A.2, where inputs and outputs are sampled at different rates. The second column gives the sampling time of the output or outputs with respect to the sampling time of the inputs. For example [8 1] means that the first output is sampled eight times slower than the inputs while the second output is sampled with the same sample time. The rest of the columns present the order of the identified system and the normalized error using the validation and identification data set. Table III shows that NucID is able to identify with rather small errors all the examples presented using the complete data set. The orders and errors are comparable to the other methods except for LR, where the order of the models is known to be high. Little performance degradation can be seen for the NucID method for all multi-rate scenarios, whereas N4SID and PEM fail for three scnearios, indicated by a star *. For the last example, the stirred tank the state space methods give very similar results to NucID, slightly outperforming it.

VII. CONCLUSIONS AND FUTURE WORKS

A system identification method, called NucID, based on nuclear norm regularization has been presented. The NucID method identifies a low order linear model from input/output data, given an upper bound on the prediction error. NucID is compared to standard identification techniques, like N4SID, prediction error minimization (PEM) and expectation conditional maximization via linear regression (LR). Diverse sets of experimental data were taken from the system identification database DaISy [9] to compare the methods among themselves. Two different scenarios of missing data in the outputs were studied. The multi-rate scenario, where the missing entries have a pattern along the outputs due to differences in the sampling times of the outputs with respect to the inputs. In the second scenario data is missing at random, e.g., when sensors fail. From the results shown in this work, we can conclude that:

- The nuclear norm regularization is a heuristic that allows one to minimize the order of the identified model. The identification problem can be posed as a convex optimization problem that yields a low order model that explains the experimental data within a given error bound.
- Normally, identifying a model form an incomplete data set involves two steps: imputing the values of missing entries in the data set according to some criteria, and then identifying a model form the “reconstructed” data set with standard system identification techniques. In contrast to this two-step approach, the NucID method involves only one step. It deals with missing data without having to make any assumptions or having to impute in some way the values of missing entries *a priori*.
- NucID can be used for system identification from complete and incomplete data sets. When data is missing at random, the advantages become clear only at high percentages of missing data. In the case of structured missing data, i.e., for multi-rate sampled-data systems, the NucID method clearly outperforms the conventional two-step procedures and is able to correctly identify a model with considerably lower sampling rates in the outputs.

REFERENCES

[1] P. Kadlec, B. Gabrys, S. Bogdan, S. Strandt, ”Data-driven Soft Sensors in the process industry”, Computers & Chemical Engineering, vol. 33, no. 4, pp. 795 - 814, 2009.

[2] B. Recht, M. Fazel and P.A. Parrilo, ”Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization”

[3] E.J. Cands, J. K. Romberg, T. Tao, ”Stable signal recovery from incomplete and inaccurate measurements”, Communications on Pure and Applied Mathematics. vol. 59, no. 8, pp. 1207-1223, 2006.

[4] D.L. Donoho. ”Compressed sensing”. IEEE Trans. Inform. Theory, vol. 52, no. 4, pp. 1289-1306, 2006.

[5] M. Fazel, H. Hindi and S. Boyd. “A rank minimization heuristic with application to minimum order system approximation”, In *Proceedings of the American Control Conference*, pp. 4734–4739, 2001

[6] Z. Liu and L. Vandenberghe. “Interior-point method for nuclear norm approximation with application to system identification”, Submitted to Mathematical Programming, 2008.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning. Data mining, inference and prediction*. Springer-Verlag, 2001.

[8] J. F. Sturm. Using SEDUMI 1.02, a Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12:625653, 1999.

[9] De Moor B.L.R. (ed.), DaISy: Database for the Identification of Systems, Department of Electrical Engineering, ESAT/SISTA, K.U.Leuven, Belgium, URL: http://homes.esat.kuleuven.be/ smc/daisy/, Oct 2008.

TABLE III

IDENTIFICATION OF EXPERIMENTAL DATA FROM DAISY DATABASE

System	norm.	NucID			N4SID			PEM			LR		
(In x Out)	T_s	n	e_V	e_I	n	e_V	e_I	n	e_V	e_I	n	e_V	e_I
Hair dryer (1x1)	1	4	0.1319	0.1387	4	0.1081	0.1222	4	0.1010	0.1219	43	0.1091	0.1171
96-006	8	4	0.1382	0.1437	*	*	*	*	*	*	43	0.1889	0.1729
Heat flow (2x1)	1	5	0.2168	0.1902	3	0.2989	0.2573	3	0.3305	0.2955	26	0.7476	0.1233
96-011	4	5	0.2881	0.2578	*	*	*	*	*	*	26	0.7545	0.1738
Heat exchanger (1x1)	1	4	0.3263	0.2130	3	0.5086	0.3742	2	0.7005	0.5245	21	0.3110	0.2059
96-002	5	5	0.3312	0.2208	*	*	*	*	*	*	21	0.3525	0.2286
Stirred tank (1x2)	[1 1]	4	0.1018	0.1511	4	0.099	0.170	3	0.143	0.1861	21	1.000	0.1487
98-002	[8 1]	3	0.1038	0.1524	4	0.0990	0.1458	3	0.0929	0.1497	*	*	*