Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets

Cristian Grossmann, Colin N. Jones, Manfred Morari

September 23, 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 simulated moving bed processes from data sets with missing entries. 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 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 compares the proposed method named NucID to the standard techniques N4SID, prediction error minimization, subspace identification and expectation conditional maximization via linear regression and a linearized first principles model. NucID is found to consistently identify systems with missing data within the imposed error tolerance, a task for 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 significantly 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. In the process and chemical industry samples might have to be collected manually and the off-line analysis can be lengthy, expensive, and wastes the valuable product as well, which makes the measurements rather sparse. 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 applied to a state-of-the-art separation process used in the pharmaceutical and fine chemical industries, namely the simulated moving bed (SMB) process. The proposed method identifies a non-parametric MIMO 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 [6], [7], [8].

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 [6], [10] 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 especially 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. The identified models are compared to a linear model derived from first principles modelling of the SMB process.

The proposed method, named NucID (because it uses the nuclear norm in the identification procedure), is found to consistently identify systems from complete data sets or data missing at random within the imposed error tolerance, a task for which the standard methods sometimes fail. In the case of structured missing data, NucID is shown to be particularly effective and to clearly outperform existing procedures. This demonstrates that NucID is an attractive tool for the identification of multi-rate sampled-data systems.

The paper is organized as follows: in the following section the SMB process is presented. The general identification problem and the identification problem with missing data are defined in Section III and V, respectively. Section IV describes the nuclear norm regularization. The methods for comparison and the results of the identification of the SMB data sets are presented in Sections VI and VII. Finally, conclusions are drawn.

II. SIMULATED MOVING BED

Simulated Moving Bed (SMB) is a continuous chromatographic process used to separate into two fractions a mixture of molecules dissolved in a fluid phase. The separation principle is based on the different affinities of the molecules in the mixture to the solid-phase which moves countercurrently to the direction of the fluid. The SMB consists of a loop of n_{col} columns where the fluid circulates in one direction (Fig. 1). The desired countercurrent flow between the two phases is achieved by switching the inlet and outlet ports in the direction of the fluid flow every t^* seconds, which

All autors are at the Automatic Control Laboratory of ETH Zurich, 8092 Zurich, Switzerland.
{grossmann,jones,morari}@control.ee.ethz.ch

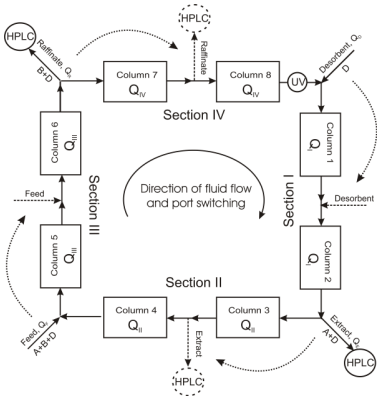

Fig. 1. Scheme of an SMB unit. The dashed lines indicate the inlet/outlet positions after the first switch. The measurements are taken by the HPLC in the extract and raffinate ports.

results in a *simulated* countercurrent movement of the solid with respect to the fluid. This periodic switching gives rise to a cyclic behavior of the process, which does not achieve a steady state with constant process variable profiles, but rather a cyclic steady state, where these profiles are repeated periodically. A detailed description of the process can be found in [2].

Economic advantages, like higher productivity and lower solvent consumption, have firmly established SMB in recent years as the state-of-the-art technology for complex separation tasks in the areas of pharmaceuticals, fine chemicals and biotechnology, especially for the purification of species characterized by low selectivities, i.e. difficult to separate, such as chiral molecules for single enantiomer drug development. [2].

The modelling, identification, optimal operation and control of SMB processes has drawn the attention of many researchers for the last decade. Several approaches have been proposed and a detailed review of these different control schemes may be found in the literature [2].

Identification of SMB models has been presented in the literature using ARX, state space models and neural networks [2]. Nevertheless it has never been considered when the data sets contain missing entries and what the least amount of measurements can be, in order to still identify a reasonable model. This is relevant for SMB separations where the measurements to be performed can be rather expensive and time consuming and one wishes to take as few measurements as possible, while on the other hand, inputs can be changed more often than the measurements can be taken.

A. SMB Virtual Plant

The data sets used in this work are produced by simulation of the SMB process with a nonlinear model. A racemic mixture of the Tröger's base enantiomers (A and B) is to be separated in a four-section SMB unit with $n_{col} = 8$ columns arranged in a 2-2-2-2 configuration as shown in Fig. 1. The dynamical model for simulation of the SMB unit is obtained by interconnecting the dynamical models of each chromatographic column. The single-column dynamics are modelled with the equilibrium dispersive model (EDM) and the adsorption behavior of both components inside the columns is described by a linear adsorption isotherm, with Henry's constants H_A and H_B. The mathematical model is completed by considering the corresponding node balances between the columns and the proper boundary and initial conditions as reported in [2], [5]. The parameters of the system under consideration are reported in Table I.

TABLE I
PHYSICAL PARAMETERS AND SMB UNIT USED FOR SIMULATION.

$Parameter$	$Value$
Henry's constants	$H_A = 5.0$
	$H_B = 1.9$
Column diameter,	$1\,cm$
Column length,	$10\,cm$
Total packing porosity	$\varepsilon = 0.68$
Theoretical plates per column	40

The internal flow rates in the four sections of the unit, Q_I, Q_{II}, Q_{III}, Q_{IV}, are used as input variables. The output measurements are the concentration levels in the extract (E) and raffinate (R) streams averaged over one cycle, $c^{ave}_{A,E}$, $c^{ave}_{B,E}$, $c^{ave}_{A,R}$, $c^{ave}_{B,R}$ as described below.

B. Output measurements

It is possible to collect samples of the outlet streams over a period of time τ and analyze them with an high performance liquid chromatography (HPLC) system. These measurements will deliver the *average* concentrations of both species, $c^{ave}_{A,j}$ and $c^{ave}_{B,j}$, in the stream j over the period of time τ

$$c^{ave}_{i,j} = \frac{\int_0^\tau c_{i,j}(t)Q_j(t)dt}{\int_0^\tau Q_j(t)dt} \tag{1}$$

for $i = A, B$. The factor $Q_j(t)$ is the flow rate of stream j, from which the sample was collected. We choose to collect samples of the extract and raffinate streams $j = E, R$, over a period of time $\tau = n_{col}\, t^*$, the cycle time. Next we present the general formulation of the identification problem considered in this work.

III. PROBLEM FORMULATION

The identification problem is first formulated for the case where no data is missing in the outputs, before being extended in Section V to the general case of missing data. A more detailed presentation has been reported in [14], but for the sake of completeness a summary is given here.

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

$$y_i(t) = \sum_{j=1}^{m} \sum_{\tau=t-r}^{t} h_{ij}(t-\tau)u_j(\tau) + v_i(t) \quad i = 1, \ldots, p \tag{2}$$

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,

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 (2) over the N samples:

$$e_N := \sum_{t=0}^{N} (y^{meas}(t) - y(t))^2 = \|Y^{meas} - Y\|_F^2 \ , \quad (3)$$

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

The FIR coefficients $h_{ij}(t)$ for $t = 0, \ldots, r$ of each of the $i \cdot j$ channels of model (2) 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}

$$\mathcal{H}_h := \begin{bmatrix} h(0) & h(1) & \cdots & h(r-n_H) \\ h(1) & h(2) & \cdots & h(r-n_H+1) \\ h(2) & h(3) & \cdots & h(r-n_H+2) \\ \vdots & \vdots & & \vdots \\ h(n_H) & h(n_H+1) & \cdots & h(r) \end{bmatrix} \quad (4)$$

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 is independent of r. The order of model (2) 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:

$$\begin{aligned} \min_h \quad & \operatorname{rank}(\mathcal{H}_h) \quad (5) \\ \text{s.t.} \quad & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned}$$

Alternatively, problem (5) can be written as

$$\min_h \|Y^{meas} - Y\|_F^2 + \alpha \operatorname{rank}(\mathcal{H}_h) \quad (6)$$

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.

IV. 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 [9] and shown in [6] to be the *convex envelope*, or the closest convex function to the rank operation:

$$\|A\|_* := \sum_{i=1}^{n} \sigma_i(A) \quad (7)$$

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 maximize sparsity in the decision vector of optimization problems, in fields ranging from statistics [11] to communications [8]. 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 (7) 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 (5) and relax the non-convex rank to a nuclear norm minimization:

$$\begin{aligned} \min_h \quad & \|\mathcal{H}_h\|_* \quad (8) \\ \text{s.t.} \quad & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned}$$

The above optimization problem can be re-cast as a semi-definite program (SDP) [9]

$$\begin{aligned} \min \quad & \mathbf{tr}(V_1) + \mathbf{tr}(V_2) \quad (9) \\ \text{s.t.} \quad & \begin{bmatrix} V_1 & \mathcal{H}_h^T \\ \mathcal{H}_h & V_2 \end{bmatrix} \succeq 0 \\ & \|Y^{meas} - Y\|_F^2 \leq \gamma \end{aligned}$$

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 (9) can therefore be posed and solved using standard SDP software (e.g., [12]).

Computational complexity: The SDP (9) 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 [10] 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 (9) with minor modification. The method [10] 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 $\|\mathcal{H}_h\|_*$.

V. SYSTEM IDENTIFICATION 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 $t = 0, \ldots, N-1$ 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 (2) to compute the matrix $Y \in \mathbb{R}^{N \times p}$, which is the *predicted* output of the model at all sample points $t = 0, \ldots, N-1$.

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}

$$e_{MD} := \|Y^{meas} - MY\|_F^2 \ . \quad (10)$$

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 (2) 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

$$\begin{aligned} \min_{h} \quad & \|\mathcal{H}_h\|_* \\ \text{s.t.} \quad & \|Y^{meas} - MY\|_F^2 \leq \gamma_{MD} \end{aligned} \tag{11}$$

A sensitivity analysis was carried out on problem (11) 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.

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.

VI. IDENTIFICATION OF SMB PROCESSES

The proposed identification method, from now on referred to as nuclear norm identification (NucID), was compared with standard toolboxes available in MATLAB. Different simulation studies were performed to check the performance of the identification method:

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.

A. Generation of identification data

The system has four inputs that are internal flow rates of the SMB unit and four outputs that are the average concentrations of the two components A and B in the two different outlet streams. The data sets used in the identification procedure and for validation were generated by computing 250 data points with the nonlinear model described in section II-A. The unit was perturbed around the point used for the linearization of the first principle model with inputs drawn from a zero-mean 1% standard deviation gaussian distribution. The different scenarios described above were simulated by dropping the outputs of the identification data set according to the corresponding approach.

B. Benchmark methods

Four 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) Subid: Estimate a state-space model [3]. (subid)
4) 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 N4SID, PEM and Subid 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 Subid.

The LR method uses a so-called expectation conditional maximization (ECM) algorithm which is a two step procedure as well [4].

These two step procedures 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.

VII. RESULTS

This section presents the identification of a dynamical model for SMB processes comparing the proposed method NucID with four standard identification tools, N4SID, PEM, Subid and LR and a linearized first principles model (FP) [5].

The models are compared throughout this paper in terms of two different criteria: the first one being the order of

the identified model and the second one the normalized prediction error on the validation set. To evaluate the first criteria, a singular value decomposition (SVD) of the Hankel matrix built from the FIR coefficients of the model at stake was computed, the order of the corresponding model was then defined as the number of singular values above 0.01% (10^{-4}) of the first value. The error on the validation set is defined as:

$$e_V := \frac{\|Y_v^{pred} - Y_v\|_F}{\|Y_v - \bar{Y}_v\|_F} \tag{12}$$

where Y_v^{pred} are the outputs predicted by the model under consideration for the validation inputs and Y_v are the outputs of the validation set. The sum of the squared errors has been normalized with the factor $\|Y_v - \bar{Y}_v\|_F$, where $\bar{Y}_v$ is the mean of the samples.

In the case of the NucID method the only tuning parameter to be chosen is the error bound γ. decreasing γ will increase the number of non-negligible singular values i.e., the order of the identified model is higher the tighter the error bound is chosen.

The complete data set contains 250 sample points out of which 125 were used for the identification procedure and the rest to validate the identified models. In the first experiment all output data was considered while for the second and the third experiments output data was dropped according to the strategy described.

1) Complete data set: In a first step, the complete data set was used to identify a dynamical model using N4SID, PEM, SubID, LR and NucID, and compare them to the FP model.

The different approaches are compared by plotting for each method the order of the identified model against the corresponding normalized validation error in Figure 2.

The FP model is mapped onto one point according to its validation error of 0.17 and order of the model of 11 since it is only one model and has no tuning parameters.

For the methods N4SID and PEM, models with fixed orders from 1 to 10 were identified on the complete data set and their normalized validation errors computed. Only the stable identified models are plotted in Figure 2. The PEM method is able to identify only models of rank 1 and 5 with very high validation errors. N4SID on the other hand is able to identify models of lower order and slightly lower validation errors.

The Subid method manages to identify models with validation errors below 0.1 for model orders between 9 and 12.

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. The validation error corresponding to this method is 0.28 but the order is of 68, hence cannot be seen in Figure 2. It is well known that LR gives a rather good fit, but with very high order models.

For the NucID method the tuning of the order is done by varying γ and the order and validation errors of seven different values of γ are plotted in Figure 2.

It is evident for the Subid and NucID methods, that there is a trade-off between the order of the identified model and the validation error. The Subid, FP and NucID models give lower validation errors than models identified with N4SID and PEM with the same order.

We can conclude that when using the complete set of data the Subid and NucID methods are able to identify dynamical models that are comparable to the FP model in terms of model order and prediction error, and that outperform N4SID, PEM and LR. Inspection of the impulse response of the Subid, NucID and FP models confirm this conclusion. The next step is to assess the impact of missing output data on the identified models with the different methods.

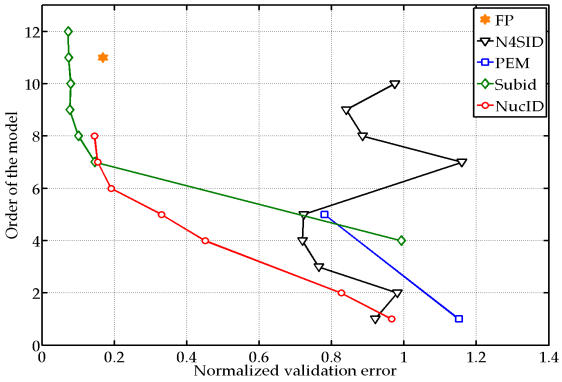

Fig. 2. Order of the identified model as a function of the normalized validation error for NucID, N4SID, PEM and FP.

2) Structured missing data: This section presents the identification of the SMB process assuming that the output data was collected at a slower sampling rate than the inputs. This is a situation that arises commonly in SMB practice when samples of the extract and raffinate stream have to be manually collected during operation for off-line analysis. It is of great interest to minimize the number of samples to be taken since the off-line analysis is lengthy and expensive and it represents a loss of the valuable product.

A scenario is presented here where samples of the extract and raffinate stream are taken and analyzed alternatingly every three cycles. Out of the four outputs measured in this approach, two of them come from the off-line analysis of the extract stream and the other two from the raffinate stream. Note that compared to the previous example, where the complete data set was used, in this case 87.5 % of the output data is missing.

The results of this scenario are presented in Figure 3. It is evident that N4SID, PEM and Subid methods suffer a severe deterioration in the quality of the model in terms of both, the order and the normalized validation error. The models identified by these methods cannot be used for predictive purposes. On the other hand, the NucID method manages to identify models with low validation errors that are comparable to the case with complete data. This example illustrates that the NucID method is able to identify low order models from data sets with structured missing data.

3) Randomly missing data: In this example, an increasing percentage of the output entries 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 $MD_{\%}$. The measurements of the extract samples (output 1 and 2) and raffinate samples (output 3 and 4) were independently and randomly dropped out according to the percentage of missing data $MD_{\%}$.
The normalized validation errors e_V for each of the methods

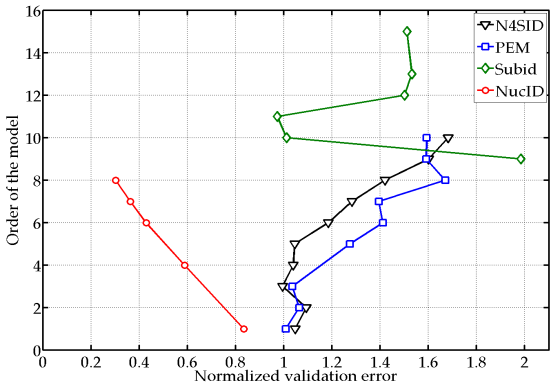

Fig. 3. Order of the identified model as a function of the normalized validation error for NucID, N4SID, PEM and Subid identified models from structured missing data.

is reported together with the order n of the identified model. For NucID, N4SID and PEM the validation error of the same order models are reported, whereas for Subid and LR the errors correspond to different order models.

N4SID's and PEM's models give validation errors above 0.7 already with $MD_{\%} = 10\%$ and remains at high validation errors throughout the increase of missing data. The Subid method shows a high sensitivity of the validation error to the percentage of missing data, which increases from 0.23 to 0.51, a considerable deterioration in the model's performance. The LR method gives low validation errors up to $MD_{\%} = 40\%$, nevertheless the orders are unreasonably high. For the last identification instance with $MD_{\%} = 60\%$ the LR method has not enough data to identify a model, indicated with a star *. NucID shows little sensitivity to the increase in the percentage of missing data and has the smallest validation errors, i.e. between 0.17 and 0.20 which in the same range as in the case of the complete data set.

TABLE II
RESULTS FOR MISSING DATA AT RANDOM

		NucID	N4SID	PEM	Subid		LR	
MD	n	e_V			n	e_V	n	e_V
10	7	0.17	0.71	1.00	9	0.23	47	0.09
20	7	0.18	0.72	0.73	9	0.26	48	0.10
30	7	0.18	0.71	0.80	9	0.28	48	0.12
40	7	0.18	0.73	0.74	9	0.37	48	0.18
50	7	0.20	0.76	0.81	9	0.50	48	0.69
60	8	0.20	0.81	0.80	9	0.51	*	*

VIII. CONCLUSIONS

A system identification method, called NucID, based on nuclear norm regularization has been presented and applied to SMB processes, modelled as a four-input four-output system. 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), subspace identification toolbox (Subid) and a first principles model (FP). Simulated data sets were taken of the system identification 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 from 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 from 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] A. Rajendran, G. Paredes, M. Mazzotti "Simulated moving bed chromatography for the separation of enantiomers", J. Chromatogr. A vol. 1216, pp. 709-738, 2009.
[3] P. VanOvershee and B. DeMoor, "Subspace identification for linear systems-theory, implementation, applications", Boston, MA: Kluwer Academic Publisher, 1996.
[4] Xiao-Li Meng and Donald B. Rubin, "Maximum Likelihood Estimation via the ECM Algorithm", Biometrika, vol. 80, no. 2, pp. 267-278, 1993.
[5] C.Grossmann, M. Amanullah, G. Erdem, M. Morari, M. Mazzotti, and M. Morbidelli. "Cycle to cycle optimizing control of simulated moving beds". AIChE Journal, vol. 54, pp. 194208, 2008.
[6] B. Recht, M. Fazel and P.A. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization", arXiv:0706.4138v1 [math.OC], 2007.
[7] 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.
[8] D.L. Donoho. "Compressed sensing". IEEE Trans. Inform. Theory, vol. 52, no. 4, pp. 1289-1306, 2006.
[9] 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
[10] Z. Liu and L. Vandenberghe. "Interior-point method for nuclear norm approximation with application to system identification", Submitted to Mathematical Programming, 2008.
[11] T. Hastie, R. Tibshirani, and J. Friedman. "The elements of statistical learning. Data mining, inference and prediction". Springer-Verlag, 2001.
[12] J. F. Sturm. Using SEDUMI 1.02, a Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12:625653, 1999.
[13] 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.
[14] C. Grossmann, C.N. Jones, M. Morari. *System identification via nuclear norm regularization from incomplete data sets*, 10th European Control Conference, Budapest, Hungary, 2009.