N2SID: Nuclear Norm Subspace Identification *

Michel Verhaegen * Anders Hansson **

* Delft Center for Systems and Control
Delft University
Delft, The Netherlands
** Division of Automatic Control
Linköping University
Linköping, Sweden

Abstract: The data equation in subspace identification relates Hankel matrices constructed from input and output data. The matrices in this equation, which are derived from the system matrices of the unknown system, contain important structural information. Such information is the low rank property of the product between the extended observability matrix and the state sequence, and the Toeplitz structure of the matrix of Markov parameters of the system in innovation form. We will show that nuclear norm subspace identification makes it possible to formulate these structural properties as constraints in a single convex multi-criteria optimization problem. This problem seeks a trade-off between a nuclear norm minimization problem to retrieve the subspace of interest and a prediction error criterion to find an optimal match between the measured and predicted output. The advantage of our method over existing unconstrained and nuclear norm constrained subspace identification methods is that the structural constraints are in general improving the estimates when dealing with short data sets, i.e. where the number of measurements is a small multiple of the system order. This advantage is demonstrated in a validation study making use of data sets from the DaSIy library.

Keywords: Subspace system identification, Nuclear norm optimization, Rank constraint, Short data batches, Alternating Direction Method of Multipliers

1. INTRODUCTION

System Identification is a key problem in a large number of scientific areas when it is of interest to model dynamic relationships among measured physical variables. In general identification is formulated as a parameter optimization problem. Many choices need to be made: the choice of parametrization of the model, the choice of cost function to be optimized, and the choices of how to solve the optimization problem, etc. For profound references we refer to Ljung [1999], Verhaegen and Verdult [2007]. Two widely used identification methods, or families of methods are so-called Prediction Error Methods (PEMs) and Subspace IDentification (SID) methods. Both approaches can be treated in the time or frequency domain, but for sake of simplicity we restrict ourselves to the time domain and to linear systems in this paper.

The core of PEMs is to parametrize the predictor or observer to generate an estimate of the output, and to then formulate an optimization problem to minimize a weighted cost function defined in terms of the difference between the measured output and the observer predicted output. This cost function is for the finite data length case generally a sample average of the trace of the covariance matrix of the prediction error. Though the PEM framework provides a vast amount of insights in studying and analyzing the estimated predictor, its main drawback is the non-convexity for general multi-variable state space models in innovation form, as considered in this paper. The lack of convexity can result in that the optimization method gets stuck in a local minimum, and thereby complicating the analysis of the numerical results, such as e.g. difficulty to distinguish between a bad model estimate due to a local minimum or due to a bad model parametrization. This parametrization is needed in the definition of the optimization problem, and thus the use of the approach can be quite complex and labor intensive for the non-expert user. However, the latter fact has been relieved somehow by software packages such as in Ljung [2007].

Motivated by the drawbacks of prediction error methods, SID methods derive approximate models rather than models that are “optimal” with respect to a chosen cost function. The approximation is based on linear algebra transformations and factorizations with structured Hankel matrices constructed form the input-output data. All existing SID methods aim to derive a low rank matrix from which key subspaces, hence the name subspace identification, are derived. The low rank approximation is in general done using a Singular Value Decomposition (SVD).
Though the early SID methods, Van Overschee and De Moor [1996], Larimore [1990], Verhaegen [1994] and many variants that were developed afterwards, did not optimize a prediction error or fitting error cost function, they turned out to be very useful in complementing PEM in a way that SID provided the needed initial estimates for PEM to start up the nonlinear parameter optimization. Despite this complementarity between the two families of system identification methods, there lacked an approach that integrated the subspace determination step of SID with the goodness of fit optimization step of PEM for identifying state space models in innovation form. Such a single integration step should enable a trade-off between the complexity of the identified state space models, as expressed e.g. by a rank constraint in the subspace determination step, with the accuracy of fitting, as expressed by a goodness of fit criterion.

The contributions made in Fazel [2002] to approximate a constraint on the rank of a matrix by minimizing its nuclear norm, lead to a number of improvements to the low rank approximation step over the classically used SVD in SID, Liu and Vandenberghe [2009b, a], Mohan and Fazel [2010], Fazel et al. [2012], Hansson et al. [2012], Liu et al. [2013]. In PEM the work on the nuclear norm lead to the integration between convex parameter optimization problems, such as in the estimation of the parameters of FIR or ARX models, with a nuclear norm constraint on the Hankel matrix of the markov parameters or similar, and this either in time or frequency domain, Smith [2012], Hjalmarsson, et. al. [2012], Wahlberg and Rojas [2013]. These parametric approaches are restricted to linear in the parameter model parametrizations and nuclear norm constraints on Hankel matrices. PEM problems for identifying state space models in innovation form that are non-linear in the parameters cannot be treated in that framework.

A missing link in consolidating the PEM and SID families is a convex solution to the multi-criteria optimization of minimizing a goodness of fit cost function on the one-step ahead prediction error of a multivariable state space model in innovation form using its observer form and minimizing the model complexity. This missing link is provided in this paper. The new SID method is denoted by Nuclear Norm Subspace IDentification (N2SID). It will be shown that two key features of N2SID make it well suited to identify accurate models of low complexity for short data length sequences. The first is that N2SID enables to add relevant structural information available about the unknown matrices in the data equation in SID that relates the Hankel matrices of input and output measurements as convex constraints. The second is that no pre-processing of the data matrices is any longer necessary through e.g. projections via instrumental variables.

The foundations for N2SID were presented in Verhaegen and Hansson [2014]. There the resulting optimization problem was solved using a Semi-Definite Programming (SDP) solver after a reformulation of the problem into an equivalent SDP problem. In addition to this the problem formulation was approximated in order to obtain a problem of manageable size for current SDP solvers. In this paper we will instead solve the problem with the Alternating Direction Method of Multipliers (ADMM). ADMM is known to be a good choice for solving regularized nuclear norm problems as the ones we are solving in this paper, Liu et al. [2013]. In order to get an efficient implementation we have customized the computations for obtaining the coefficient matrix of the normal equations associated with our formulation using Fast Fourier Transformation (FFT) techniques. This is the key to obtain an efficient implementation.

The paper is organized as follows. In Section 2 the identification problem for identifying a multi-variable state space model in a subspace context while taking a prediction error cost function into consideration is presented. The problem formulation does however not require a parametrization of the system matrices of the state space model as is classically done in prediction error methods for parametric input-output models, Ljung [1999]. The key data equation in the analysis and description of subspace identification methods is presented in Section 3. Here we also highlight the important structural properties of the matrices of this equation: the low rank property of the matrix that is the product of the extended observability matrix and the state sequence of the underlying state space model given in innovation form, and the lower triangular Toeplitz structure of the model Markov parameters. A convex relaxation is presented in Section 4 to take these structural constraints into consideration. In Section 5 we explain how to obtain an efficient ADMM code. The results of the performances are illustrated in Section 6 in a comparison study of N2SID with two other SID methods, N4SID and the recent Nuclear Norm based SID methods presented in Liu et al. [2013] and with PEM. The comparison study is based on the real-life data sets in the DaSy database, De Moor et al. [1997]. Finally, we end this paper with some concluding remarks.

### 1.1 Notations

We introduce the Matlab-like notation that for a vector or matrix $X \in \mathbb{R}^{M \times N}$ (or $N \times M$) it holds that $X_{m:n,p:q}$ is the sub-matrix of $X$ with rows $m$ through $n$ and columns $p$ through $q$. If one of the dimensions of the matrix is $n$, then $1:n$ can be abbreviated as just $n$. Moreover, with $X_{m:-1:n,p:q}$ is meant the sub-matrix of $X$ obtained by taking rows $m$ through $n$ in reverse order, where $m$ is greater than or equal to $n$. Similar notation may be used for the columns.

### 2. THE SUBSPACE IDENTIFICATION PROBLEM

In system identification a challenging problem is to identify Linear Time Invariant (LTI) systems with multiple inputs and multiple outputs using short length data sequences. Taking process and measurement noise into consideration, a general state space model for LTI systems can be given in so-called innovation form, Verhaegen and Verdult [2007]:

$$
\begin{align*}
\begin{cases}
x(k+1) &= Ax(k) + Bu(k) + Ke(k) \\
y(k) &= Cx(k) + Du(k) + e(k)
\end{cases}
\end{align*}
$$

with $x(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}^p$, $u(k) \in \mathbb{R}^m$ and $e(k)$ a zero-mean white noise sequence.

We will consider mainly the scenario in which both the input and output data sequences are given. Also the case of output only identification is briefly considered in
Remark 1, making the approach outlined in this paper a general and new framework to identify linear dynamical systems.

Since we are interested in short data sets no requirement on consistency is included in the following problem formulation.

**Problem Formulation:** Given the input-output (i/o) data batches \( \{u(k), y(k)\}_{k=1}^{N} \), with \( N > n \) and assumed to be retrieved from an identification experiment with a system belonging to the class of LTI systems as represented by (1), the problem is to determine approximate system matrices \( (A_T, B_T, C_T, D_T, K_T) \) that define the \( n \)-th order observer of "low" complexity:

\[
\begin{aligned}
\dot{x}_T(k+1) &= A_T x_T(k) + B_T u_v(k) + K_T (y_v(k) - C_T x_T(k)) \\
\dot{y}_v(k) &= C_T x_T(k) + Du_v(k)
\end{aligned}
\]

such that the approximated output \( \hat{y}_v(k) \) is "close" to the measured output \( y_v(k) \) of the validation pair \( \{u_v(k), y_v(k)\}_{k=1}^{N_v} \) as expressed by a small value of the cost function,

\[
\frac{1}{N_v} \sum_{k=1}^{N_v} \|y_v(k) - \hat{y}_v(k)\|^2 \quad (3)
\]

The quantitative notions like "low" and "close approximation" will be made more precise in the new N2SID solution toward this problem.

A key starting point in the formulation of subspace methods is the relation between structured Hankel matrices constructed from the i/o data. This relationship will as defined in Verhaegen and Verdult [2007] be the data equation. It will be presented in the next section. There we will also highlight briefly how existing subspace methods have missed to take important structural information about the matrices in this equation into account from the first step of the solution procedure.

### 3. THE DATA EQUATION AND ITS STRUCTURE

Let the LTI model (1) be represented in its so-called observer form:

\[
\begin{aligned}
x(k+1) &= (A - KC)x(k) + (B - KD)u(k) + Ky(k) \\
y(k) &= Cx(k) + Du(k) + e(k)
\end{aligned}
\]

We will denote this model compactly as:

\[
\begin{aligned}
x(k+1) &= Ax(k) + Bu(k) + Ky(k) \\
y(k) &= Cx(k) + Du(k) + e(k)
\end{aligned}
\]

with \( A \) the observer system matrix \((A - KC)\) and \( B \) equal to \((B - KD)\). Though this property will not be used in the sequel, the matrix \( A \) can be assumed to be asymptotically stable.

For the construction of the data equation, we store the measured i/o data in block-Hankel matrices. For fixed \( N \) assumed to be larger then the order \( n \) of the underlying system, the definition of the number of block-rows fully defines the size of these Hankel matrices. Let this dimensioning parameter be denoted by \( s \), and let \( s > n \), then the Hankel matrix of the input is defined as:

\[
U_s = \begin{bmatrix} u(1) & u(2) & \cdots & u(N-s+1) \\ u(2) & u(3) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ u(s) & u(s+1) & \cdots & u(N) \end{bmatrix} \quad (6)
\]

The Hankel matrices from the output \( y(k) \) and the innovation \( e(k) \) are defined similarly and denoted by \( Y_s \) and \( E_s \), respectively. The relationship between these Hankel matrices, that readily follows from the linear model equations in (5), require the definition of the following structured matrices. First we define the extended observability matrix \( O_s \):

\[
O_s^T = \begin{bmatrix} C^T & A^T C^T & \cdots & A^{T^{s-1}} C^T \end{bmatrix} \quad (7)
\]

Second, we define a Toeplitz matrix from the quadruple of systems matrices \( \{A, B, C, D\} \) as:

\[
T_{u,s} = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ CA^{s-2}B & \cdots & D \end{bmatrix} \quad (8)
\]

and in the same way we define a Toeplitz matrix \( T_{y,s} \) from the quadruple \( \{A, K, C, 0\} \). Finally, let the state sequence be stored as:

\[
X = [x(1) \ x(2) \ \cdots \ x(N-s+1)] \quad (9)
\]

Then the data equation compactly reads:

\[
Y_s = O_s X + T_{u,s} U_s + T_{y,s} Y_s + E_s \quad (10)
\]

This equation is a simple linear matrix equation that highlights the challenges in subspace identification, which is to approximate from the given Hankel matrices \( Y_s \) and \( U_s \) the column space of the observability matrix and/or that of the state sequence.

The equation is highly structured. In this paper we focus on the following key structural properties about the unknown matrices in (10):

1. The matrix product \( O_s X \) is low rank since \( s > n \).
2. The matrices \( T_{u,s} \) and \( T_{y,s} \) are block-Toeplitz.
3. The matrix \( E_s \) is block-Hankel.

In all existing subspace identification methods, these key structural matrix properties are not used in the first step of the algorithmic solution. Some pre-processing step of the data Hankel matrices is usually performed, followed by a low rank factorization. For example, the first step of the SID method in Chiuso [2007] is the calculation of the parameters of a high order VARX model prior to the subspace revealing step with a classical SVD. Other classes of SID methods, such as the original contributions by Van Overschee and De Moor [1996], Verhaegen [1994] make use of instrumental variables and projections to transform the original data. It is well known that such projections lead to "a loss of information" in the original data set. Such loss may be detrimental when dealing with small length data sets. A way to improve the estimates, especially for such cases, is the use of structural knowledge to constraint the model identification. For identifying the deterministic part of the innovation model Liu et al. [2013] recently replaced the SVD in the classical SID methods Van Overschee and De Moor [1996], Larimore [1990], Verhaegen [1994] with a nuclear norm optimization, to
try to enforce the low rank property to an already pre-processed matrix.

Though statistical consistency generally holds for the existing subspace identification schemes, they refrain from using key structural matrix properties in the data equation in the first step of the algorithm. Therefore in the next section we will revise subspace identification and formulate the new N2SID approach.

4. N2SID

4.1 Pareto optimal Subspace Identification

When assuming the optimal observer given, the quantity \( \hat{y}(k) \) is the minimum variance prediction of the output and equal to \( y(k) - e(k) \). Let the Hankel matrix \( \hat{Y}_s \) be defined from this sequence \( \hat{y}(k) \) as we defined \( Y_s \) from \( y(k) \). Then the data equation (10) can be reformulated into:

\[
\hat{Y}_s = O_s X + T_{u,s} U_s + T_{y,s} Y_s \quad (11)
\]

Let \( T_{p,m} \) denote the class of lower triangular block-Toeplitz matrices with block entries \( p \times m \) matrices and let \( H_p \) denote the class of block-Hankel matrices with block entries of \( p \) column vectors. Then the three key structural properties listed in Section 3 are taken into account in the following problem formulation:

\[
\min_{Y_s \in H_p, T_{u,s} \in T_{p,m}, T_{y,s} \in T_{p,p}} \text{rank}(\hat{Y}_s - T_{u,s} U_s - T_{y,s} Y_s) \quad (12)
\]

and

\[
\min \text{Tr} \left[ E \left( (y(k) - \hat{y}(k)) (y(k) - \hat{y}(k))^T \right) \right], \quad \text{where } E \text{ denotes the expectation operator.}
\]

This optimization problem seeks for the Pareto optimal solution with respect to the two cost functions \( \text{rank}(\hat{Y}_s - T_{u,s} U_s - T_{y,s} Y_s) \) and \( \text{Tr} \left[ E \left( (y(k) - \hat{y}(k)) (y(k) - \hat{y}(k))^T \right) \right] \). This optimization is however not tractable. For that purpose we will develop in the next subsection a convex relaxation. This will make it possible to obtain all Pareto optimal solutions using scalarization.

4.2 A convex relaxation

A convex relaxation of the NP hard problem formulation in (12) will now be developed. The original problem is reformulated in two ways. First, the rank operator is substituted by the nuclear norm. The nuclear norm of a matrix \( X \) denoted by \( \| X \|_n \) is defined as the sum of the singular values of the matrix \( X \). It is also known as the trace norm, the Ky Fan norm or the Schatten norm, Liu and Vandenberghe [2010]. This is known to be a good approximation of the rank operator when it is to be minimized, Fazel et al. [2001], Fazel [2002]. Second, the minimum variance criterion is substituted by the following sample average of \( \text{Tr} \left[ E \left( (y(k) - \hat{y}(k)) (y(k) - \hat{y}(k))^T \right) \right] \):

\[
\frac{1}{N} \sum_{k=1}^{N} \| y(k) - \hat{y}(k) \|^2_2
\]

By introducing a scalarization, or regularization, parameter \( \lambda \in [0, \infty) \) all Pareto optimal solutions of the convex reformulation of the N2SID problem can be formulated in one line:

\[
\min_{\hat{Y}_s \in H_p, T_{u,s} \in T_{p,m}, T_{y,s} \in T_{p,p}} \| \hat{Y}_s - T_{u,s} U_s - T_{y,s} Y_s \|_* + \lambda \frac{1}{N} \sum_{k=1}^{N} \| y(k) - \hat{y}(k) \|^2_2 
\]

\[
(13)
\]

Remark 1. The method encompasses in a straightforward manner the identification problems with output data only. In that case the convex relaxed problem formulation reads:

\[
\min_{\hat{Y}_s \in H_p, T_{y,s} \in T_{p,p}} \| \hat{Y}_s - T_{y,s} Y_s \|_* + \lambda \frac{1}{N} \sum_{k=1}^{N} \| y(k) - \hat{y}(k) \|^2_2 
\]

\[
(14)
\]

It is well-known that this problem can be recast as a Semidefinite Programming (SDP) problem, Fazel et al. [2001], Fazel [2002], and hence it can be solved in polynomial time with standard SDP solvers. The reformulation, however, introduces additional matrix variables of dimension \( N \times N \), unless the problem is not further approximated using randomization techniques as in Verhaegen and Hansson [2014]. In section 5 we will present an alternative exact method using ADMM inspired by its successful application in Liu et al. [2013].

4.3 Calculation of the system matrices

The convex-optimization problem (13) yields the estimates of the quantities \( \hat{Y}_s, T_{u,s} \) and \( T_{y,s} \). Since the outcome depends on the regularization parameter \( \lambda \), we denote these estimates as \( \hat{Y}_s^\lambda, T_{u,s}^\lambda \) and \( T_{y,s}^\lambda \) respectively. The determination of the system matrices starts with an SVD of the “low rank” approximated matrix as follows:

\[
\hat{Y}_s^\lambda - T_{u,s}^\lambda U_s - T_{y,s}^\lambda Y_s = \begin{bmatrix} U_n(\lambda) & \ast \end{bmatrix} \frac{\Sigma(\lambda)}{0} \begin{bmatrix} V_n^T(\lambda) & \ast \end{bmatrix}
\]

\[
(15)
\]

where \( n \) is an integer denoting the \( n \) largest singular values and the notation \( \ast \) denotes a compatible matrix not of interest here. The selection of \( n \) is outlined in the algorithmic description given next.

The algorithm requires in addition to the input-output data sequences the user to specify the parameter \( s \) to fix the number of block rows in the block-Hankel matrices \( U_s \) and \( Y_s \) and an interval for the parameter \( \lambda \) denoted by \( \Lambda = [\lambda_{\min}, \lambda_{\max}] \). As for the implementation described in Liu et al. [2013], which we will refer to as \texttt{N2SID}, the identification data set could be partitioned in two parts. The first part is referred to as the \texttt{ide-1} part of the identification data set and the remaining part of the identification data set is referred to as the \texttt{ide-2} part. This splitting of the data set was recommended in Liu et al. [2013] to avoid overfitting. In the \texttt{N2SID} algorithm three variants can be substituted in the algorithmic block ‘compute \( M_j(\lambda) \)’ for \( j = 1, 2, 3 \). This algorithmic block performs the actual calculation of the one-step ahead predictor and the three variants are summarized after the description of the core part of \texttt{N2SID}. 

**N2SID algorithm:**

Grid the interval $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ in $N$ different points, e.g. using the Matlab notation $\Lambda = \text{logspace}(\log(\lambda_{\min}), \log(\lambda_{\max}), L)$

for $i = 1:L$,

Solve (13) for $\lambda = \Lambda(i)$ and data set ide-1. 

Compute the SVD as in (15) for $\lambda = \Lambda(i)$.

Select Select the model order $\hat{n}$ form the singular values in (15). This can be done manually by the user or automatically. Such automatic selection can be done as in the N4SID implementation in Ljung [2007] as highlighted in Liu et al. [2013]; order the singular values in (15) in descending order, then select that index of the singular value that in logrithm is closest to the logarithmic mean of the maximum and minimum singular values in (15).

Compute system matrices $\{\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T, \hat{K}_T\}$ according to the procedure 'Compute $\mathcal{M}_j(\lambda)$' for $j = 1, 2, 3$ and $\lambda = \Lambda(i)$.

Using the estimated system matrices $\{\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T\}$, and the validation data in ide-2, compute the simulated output $\hat{y}(k, \lambda)$ as,

$$
\hat{x}_T(k + 1) = \hat{A}_T \hat{x}_T(k) + \hat{B}_T u(k)
$$

$$
\hat{y}(k, \lambda) = \hat{C}_T \hat{x}_T(k) + \hat{D}_T u(k)
$$

and evaluate the cost function,

$$
J(\lambda) = \sum_{i=1}^{N} \|y(k) - \hat{y}(k, \lambda)\|^2_2
$$

end

Select $\mathcal{M}_j(\lambda_{\text{opt}})$ with $\lambda_{\text{opt}}$ given as:

$$
\lambda_{\text{opt}} = \min_{\lambda \in \Lambda} J(\lambda)
$$

The subsequent three ways to compute the model are summarized as:

Compute $\mathcal{M}_1(\lambda)$:

**STEP 1:** From the SVD in (15), and the selected model order $\hat{n}$, the pair $\hat{A}_T, \hat{C}_T$ is derived from the matrix $\hat{U}_n$ as done in classical SID methods by considering $\hat{U}_n$ to be an approximation of the extended observability matrix $\mathcal{O}_s$, see e.g. Verhaegen and Verdult [2007].

**STEP 2:** With $\hat{U}_n$ and the estimated matrix $\hat{T}_{y,s}$ we exploit that the latter matrix approximates the block-Toeplitz matrix $\hat{T}_{u,s}$ to estimate the observer gain $\hat{K}_T$ via the solution of a standard linear least squares problem. This estimate of the observer gain is used to estimate the system matrix $\hat{A}_T$ as:

$$
\hat{A}_T = \hat{A}_T + \hat{K}_T \hat{C}_T \quad (17)
$$

**STEP 3:** Let the approximation of the observer be denoted as:

$$
\hat{x}_T(k + 1) = \hat{A}_T \hat{x}_T(k) + \hat{B}_T u(k) + \hat{K}_T y(k)
$$

$$
\hat{y}(k) = \hat{C}_T \hat{x}_T(k) + \hat{D}_T u(k)
$$

Then the estimation of the pair $\hat{B}_T, \hat{D}$ and the initial conditions of the above observer can again be done via a linear least squares problem as outlined in Verhaegen and Verdult [2007] by minimizing the RMS value of the prediction error $y(k) - \hat{y}(k)$ determined from the identification data in ide-1. The estimated input matrix $\hat{B}_T$ is then determined as:

$$
\hat{B}_T = \hat{B}_T + \hat{K}_T \hat{D} \quad (19)
$$

Compute $\mathcal{M}_2(\lambda)$:

**STEP 1:** as in Compute $\mathcal{M}_1(\lambda)$.

**STEP 2:** Derive an estimate of the state sequence of the observer (18) from the SVD (15), where for the sake of compactness again the system symbol $\hat{x}_T(k)$ will be used,

$$
[\hat{x}_T(1) \hat{x}_T(2) \cdots \hat{x}_T(N - s + 1)] \approx V_n^T(\lambda)
$$

Using the singular values this approximation could also be scaled as $\sqrt{\sum_{\lambda} V_n^T(\lambda)}$.

**STEP 3:** Knowledge of the estimated state sequence of the observer (18) turns the estimation of the system matrices $\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}_T, \hat{K}_T$ and the observer initial conditions into linear least squares problem. The estimated pair $(\hat{A}_T, \hat{B}_T)$ can be computed from this quintuple as outlined in (17) and (19), respectively.

Compute $\mathcal{M}_3(\lambda)$:

**STEP 1 and 2:** as in Compute $\mathcal{M}_1(\lambda)$.

**STEP 3:** With $\hat{U}_n$ and the estimated Markov parameters in $\hat{T}_{u,s}$, we could similarly to estimating the Kalman gain also estimate the pair $\hat{B}_T, \hat{D}$ via a linear least squares problem. The matrix $\hat{B}_T$ can be estimated from $\hat{B}_T$ as outlined in (19).

In the experiments reported in Section 6 use will be made of N2SID Algorithm with the model computation block Compute $\mathcal{M}_1(\lambda)$. It turned out that the N2SID algorithm is much less sensitive to overparametrization compared as compared to WNNOpt. For that reason we will use the whole identification data set in all steps of the N2SID algorithm for the experiments reported in Section 6, i.e. ide-1 and ide-2 are identical and equal to the identification data set.

5. ADMM

The problem we like to solve is exactly of the form in (20) in Liu et al. [2013], i.e.

$$
\min_x \| A(x) + A_0 \|_\infty + \frac{1}{2} (x - a)^T H (x - a) \quad (20)
$$

for some linear operator $A(x)$ and some positive semidefinite matrix $H$. In the above mentioned reference the linear operator is a Hankel matrix operator, and this structure is used to tailor the ADMM code to run efficiently. Essentially the key is to be able to compute the coefficient matrix related to the normal equations of the linear operator in an efficient way using FFT. This matrix $M$ is defined via

$$
A_{\text{adj}}(A(x)) = M x, \forall x
$$

where $A_{\text{adj}}(\cdot)$ is the adjoint operator of $A(\cdot)$. Similar techniques have been used for Toeplitz operators in Roh and Vandenberge [2006], and are closely related to techniques for exploiting Toeplitz structure in linear systems of equations, Golub and van Loan [1996]. For our problem the linear operator consists of a sum of Hankel and Toeplitz operators, and we will show how FFT techniques can be used also for this operator.
5.1 Circulant, Toeplitz and Hankel Matrices

We define the circulant matrix operator $C^n : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ of a vector $x \in \mathbb{R}^n$ via

$$C^n(x) = \begin{bmatrix} x_1 & x_n & \cdots & x_3 & x_2 \\ x_2 & x_1 & x_n & \cdots & x_3 \\ \vdots & x_2 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ x_{n-1} & \cdots & \cdots & x_1 & x_n \\ x_n & x_{n-1} & \cdots & x_2 & x_1 \end{bmatrix} \quad (21)$$

We also define the Hankel matrix operator $H^{(m,n)} : \mathbb{R}^{m+n-1} \to \mathbb{R}^{m \times n}$ of a vector $x \in \mathbb{R}^{m+n-1}$ via

$$H^{(m,n)}(x) = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_m & \cdots & \cdots & x_{m+n-1} \end{bmatrix} \quad (22)$$

For a vector $x \in \mathbb{R}^{m+n-1}$ it holds that $H^{(m,n)}(x) = C^{m+n-1}_{n-1,1}(x)$, i.e. the Hankel operator is the lower left corner of the circulant operator where the columns are taken in reverse order. We also define the Toeplitz operator $T^n : \mathbb{R}^{2n} \to \mathbb{R}^{n \times n}$ of a vector $x \in \mathbb{R}^{2n}$ via

$$T^n(x) = \begin{bmatrix} x_n & x_{n-1} & \cdots & x_1 \\ x_{n+1} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n-1} & \cdots & \cdots & x_n \end{bmatrix} \quad (23)$$

We realize that $T^n(x) = H^{(n,n)}_{n-1,1}(x)$, i.e. a Toeplitz operator can be obtained from a square Hankel operator by taking the columns in reverse order. We are finally interested in upper triangular Toeplitz operators with and without zeros on the diagonal, and we remark that these are easily obtained from the normal Toeplitz operator by replacing $x$ with $\begin{bmatrix} x^T \\ 0 \end{bmatrix}$, which will be upper triangular with a non-zero diagonal if $x \in \mathbb{R}^n$ and with a zero diagonal if $x \in \mathbb{R}^{n-1}$.

5.2 The Fourier Transform and Hankel Matrices

It is well-known, Golub and van Loan [1996], that if we let $F^n \in \mathbb{C}^{n \times n}$ be the discrete Fourier transform matrix of dimension $n$, then the circulant matrix can be expressed as

$$C^n(x) = \frac{1}{N} F^n H \text{diag}(F^n x) F^n \quad (24)$$

From this we immediately obtain that the Hankel matrix can be expressed as

$$H^{(m,n)}(x) = \frac{1}{N} H F \text{diag}(F^N x) G \quad (25)$$

where $N = m + n - 1$, $F = F^N$, $G = F_{n-1:1}$ and $H = F_{n:m+n-1}$. This expression will make it easy for us to represent the adjoint of the Hankel operator. It is straightforward to verify that the adjoint $H^{(m,n)}_{\text{adj}}(Z) : \mathbb{R}^{m \times n} \to \mathbb{R}^{n \times m-1}$ is given by

$$H^{(m,n)}_{\text{adj}}(Z) = \frac{1}{N} F^H \text{diag}(HZG^H)$$

Notice that we are abusing the operator diag(·). In case the argument is a vector the operator produces a diagonal matrix with the vector on the diagonal, and in case the argument is a square matrix, the operator produces a vector with the components equal to the diagonal of the matrix.

5.3 The Linear Operator $A$

We will now present the linear operator that we are interested in for the SISO case: $A : \mathbb{R}^N \times \mathbb{R}^s \times \mathbb{R}^{s-1} \to \mathbb{R}^{s \times n}$, where

$$A(x) = H^{(s,n)}(y) + T^s \begin{bmatrix} u_{s-1:1} \\ 0 \end{bmatrix}^T V$$

$$+ T^s \begin{bmatrix} w_{s-1:1} \\ 0 \end{bmatrix}^T W$$

where $n = N - s + 1$, $x = (y, v, w)$ with $y \in \mathbb{R}^N$, $v \in \mathbb{R}^s$, $w \in \mathbb{R}^{s-1}$, $V \in \mathbb{R}^{s \times N}$ and $W \in \mathbb{R}^{s \times s}$. By taking $V = -U_s$ and $W = -Y_s$ we obtain the linear operator for N2SID. Then, $v$ and $w$ are the first columns of the Toeplitz matrices $T_{u,s}$ and $T_{y,s}$, respectively. We can express $A(x)$ in terms of Hankel operators as:

$$A(x) = H^{(s,n)}(\hat{y}) + H^{(s,s)}_{1:1}(u_{s-1:1})^T V$$

$$+ H^{(s,s)}_{1:1}(w_{s-1:1})^T W$$

The adjoint of this operator can be expressed in terms of the adjoint of the Hankel operator as

$$A_{\text{adj}}(Z) = \begin{bmatrix} H^{(s,n)}_{\text{adj}}(Z) \\ H^{(s,s)}_{\text{adj},s-1:1}(W Z^T_{s-1:1}) \\ H^{(s,s)}_{\text{adj},1:s-1:1}(W Z^T_{1:s-1:1}) \end{bmatrix}$$

5.4 Forming the Coefficient Matrix

A key matrix in the ADDM algorithm is the matrix $M$ defined via

$$A_{\text{adj}}(A(x)) = Mx, \quad \forall x$$

We will now show how this matrix can be formed efficiently using the Fast Fourier Transform (FFT). We partition the matrix as

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{12}^T & M_{22} & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{bmatrix}$$

where the partition is done to conform with the partition $x = (y, v, w)$. It is then clear that $M_{11}$ is defined via

$$H^{(s,n)}_{\text{adj}}(H^{(s,n)}(y)) = M_{11} y, \quad \forall y$$

The left hand side can be expressed as

$$\frac{1}{N^2} F^H \text{diag} \left( HH^H \text{diag} (Fy) GG^H \right)$$

From the identity

$$\text{diag}(A \text{ diag}(x) B) = (A \odot B^T) x$$

where $\odot$ denotes the Hadamard product of matrices, it follows that

$$M_{11} = \frac{1}{N^2} F^H \left( (HH^H) \odot \left( GG^H \right) \right) F$$
The efficient way to form $M_{11}$ is to first compute $F_i G_i$ and $H_i$ using an FFT algorithm, and then to form the matrix

$$X = (HH^H) \odot (GG^H)$$

After this one should apply the inverse FFT algorithm to $X^T$, and then to the transpose of the resulting matrix once more the inverse FFT algorithm.

The expressions for the other blocks of the matrix $M$ can be derived in a similar way, and they are given by:

$$M_{12} = \frac{1}{N_2} F_i^H \left( (H_{i:s-1:1} H^H) \odot (GV^i G_i^H) \right) F_{s-1:1}$$

$$M_{13} = \frac{1}{N_2} F_i^H \left( (H_{i,s-1:1} H^H) \odot (GW^i G_i^H) \right) F_{s-1:1}$$

$$M_{22} = \frac{1}{N_2} F_i^H \left( (H_{i:s-1:1} H^H) \odot (GG^H) \right) F_{s-1:1}$$

$$M_{23} = \frac{1}{N_2} F_i^H \left( (HW_i H^H) \odot (GG^H) \right) F_{s-1:1}$$

$$M_{33} = \frac{1}{N_2} F_i^H \left( (HW_{i:s-1:1} H^H) \odot (GG^H) \right) F_{s-1:1}$$

Notice that for the last three blocks the matrices $F_i$, $G_i$, and $H_i$ are defined via a discrete Fourier transform matrix of order $\kappa = 2s - 1$.

5.5 MIMO Systems

So far we have only discussed SISO systems. For a general $p \times m$ system we may write the linear operator $A : \mathbb{R}^{pN} \times \mathbb{R}^{pm} \times \mathbb{R}^{p(s-1)} \rightarrow \mathbb{R}^{p \times n}$ as:

$$A(x) = \sum_{i=1}^{p} A_i(x_i) \odot e_i$$

where

$$A_i(x_i) = \mathcal{H}^{(s,n)}(\gamma_i) + \sum_{j=1}^{m} \mathcal{H}^{(s,s)}_{i:s-1:1} \begin{bmatrix} w_{i,j;1:1} \\ 0 \end{bmatrix}^T V_j$$

$$+ \sum_{j=1}^{p} \mathcal{H}^{(s,s)}_{i:s-1:1} \begin{bmatrix} w_{i,j;1:1} \\ 0 \end{bmatrix}^T W_j$$

where $V_j = -U_j^T$ and $W_j = -Y_j^T$ are Hankel matrices defined from $u_j$ and $y_j$, i.e. from the $j$th inputs and outputs, respectively. Hence we may interpret each term $A_i(x_i)$ as defining a MISO system in the sense that each predicted output can be written as a linear combination of all the inputs and outputs. If we write the adjoint variable $Z$ in a similar way as $Z = \sum_{i=1}^{p} Z_i \odot e_i$, it follows that the adjoint operator is given by $A_{adj}(Z) = (A_{adj,1}(Z_1), \ldots, A_{adj,p}(Z_p))$. Hence the matrix $M$ will now be blockdiagonal with blocks defined from the identity

$$A_{adj,i}(A_i(x_i)) = M_{i} x_i, \quad \forall x_i, i = 1, \ldots, p$$

It is not difficult to realize that the operators $A_{adj,i}(Z_i)$ will be given by

$$A_{adj,i}(Z_i) = \begin{bmatrix} \mathcal{H}^{(s,s)}_{i,s-1:1}(V_1 Z_i; c_{i-1:1}) \\ \vdots \\ \mathcal{H}^{(s,s)}_{i,s-1:1}(V_m Z_i; c_{i-1:1}) \\ \mathcal{H}^{(s,s)}_{i,s-1:1}(W_1 Z_i; c_{i-1:1}) \\ \vdots \\ \mathcal{H}^{(s,s)}_{i,s-1:1}(W_p Z_i; c_{i-1:1}) \end{bmatrix}$$

Hence each of the blocks $M_i$ will have a similar structure as the $M$ matrix for the SISO system. However, the sub-blocks $M_{12}$, $M_{13}$, $M_{22}$, $M_{23}$ and $M_{33}$ will have sub-blocks themselves reflecting that fact that there are several inputs and outputs. $M_{11}$ will be the same as $M_i$ for the SISO case for all $i$. Below are formulas given for sub-blocks of each of the other matrices

$$M_{12j} = \frac{1}{N_2} F_i^H \left( (H_{i:s-1:1} H^H) \odot (GV^i G_i^H) \right) F_{s-1:1}$$

$$M_{13j} = \frac{1}{N_2} F_i^H \left( (H_{i,s-1:1} H^H) \odot (GW^i G_i^H) \right) F_{s-1:1}$$

$$M_{22j} = \frac{1}{N_2} F_i^H \left( (H_{i:s-1:1} H^H) \odot (GG^H) \right) F_{s-1:1}$$

$$M_{23j} = \frac{1}{N_2} F_i^H \left( (HW_{i:s-1:1} H^H) \odot (GG^H) \right) F_{s-1:1}$$

$$M_{33j} = \frac{1}{N_2} F_i^H \left( (HW_{i:s-1:1} H^H) \odot (GG^H) \right) F_{s-1:1}$$

It is interesting to notice that these formulas do not depend on index $i$ neither. This means that all $M_i$ are the same.

6. VALIDATION STUDY

In this section we report results on numerical experiments using real-life data sets. We will make use of some representative data sets from the DaISy collection, De Moor et al. [1997]. For preliminary test with the new N2SID method based on academic examples, we refer to Verhaegen and Hansson [2014].

The numerical results reported in Subsection 6.3 were performed with Matlab. The implementations have been carried out in MATLAB R2013b running on an Intel Core i7 CPU M 250 2 GHz with 8 GB of RAM.

6.1 Data selection and pre-processing

From the DaISy collection, De Moor et al. [1997], five representative data sets were selected. These sets contain SISO, SIMO, MISO and MIMO systems. Information about the selected data sets is provided in Table 1. In order to evaluate the performances for small length data sets, data sets of increasing length are considered. The data length is indicated by $N_{ide}$ in Table 2 for each data set of Table 1. To test the sensitivity of the identification methods with respect to the length of the identification data set, $N_{ide}$ is increased from a small number, as compared to the total number of samples available, in a way as indicated in Table 2. From each identification and validation data set the offset is removed. Data set 2 from the continuous stirred tank reactor is scaled in such a way that both outputs have about the same numerical range. This is
achieved by scaling the detrended versions of these outputs such that the maximum value of each output equals 1.

Since many of the data sets contain poorly excited data at their beginning, the first \( \text{del} \) samples are discarded from each data set. The actual value of \( \text{del} \) for each data set is listed in Table 2. Finally each identified model is validated for each test case on the same validation data set. These validation data sets contain the \( N_{\text{val}} \) samples following the sample with index \( \max(N_{\text{ide}}) + 1 \). The value of \( N_{\text{val}} \) is listed in Table 2.

### 6.2 Compared Identification methods

Three SID methods are compared in the tests. Their key user selection parameters are listed in Table 3. One of the key user selection parameters of the SID methods is the number of block rows \( s \) of the Hankel data matrices. In methods like \( \text{N4SID} \) or \( \text{WNNopt} \) of Table 3 a distinction could be made in the number of block rows of so-called future and past Hankel matrices. Such differentiation is not necessary for the \( \text{N2SID} \) algorithm. Since such differentiation is still an open research problem we opted in this simulation study to take the number of block rows of the future and past Hankel matrices in \( \text{N4SID} \) or \( \text{WNNopt} \) equal to the number of block rows in the \( \text{N2SID} \) algorithm. Table 3 also lists the interval of the regularization parameter \( \lambda \) to be specified for the Nuclear Norm based methods.

For \( \text{N4SID} \) we further used the default settings except that the Kalman filter gain is not estimated, a guaranteed stable simulation model is identified, no input delays are estimated and these are fixed to zero, and finally no covariance estimates are determined. The order selection is done with \( \text{N4SID} \) using the option `best', Ljung [2007]. This results in a similar automatic choice as we have implemented for \( \text{N2SID} \).

For \( \text{WNNopt} \) in Liu et al. [2013] the weighting according to the CVA method is used. Also here no Kalman gain is estimated and the input delay is set to zero. As indicated in Liu et al. [2013] an 'identification' and a 'validation' data set is needed to perform the selection of the regularization parameter \( \lambda \) in order to avoid overfitting. In this paper both data sets are retrieved from the identification data set of length \( N_{\text{ide}} \) by splitting it into two almost equal parts differing in length by at most one sample.

For \( \text{N2SID} \) we used the ADMM algorithm presented in Liu et al. [2013], where we have provided our own routines for computing \( M \) as explained in section 5. As explained in Liu et al. [2013] we also make use of simultaneous diagonalization of \( M \) and the positive semidefinite matrix \( H \) in (20) in order not to have to make different factorizations for each value of the regularization parameter \( \lambda \). The maximum number of iterations in the ADMM algorithm have been set to 200, the absolute and relative solution accuracy tolerances have been set to \( 10^{-6} \), and \( 10^{-3} \), respectively. The parameters used to update the penalty parameter have been set to \( \tau = 2 \) and \( \mu = 10 \). We label our \( \text{N2SID Algorithm with N2SID} \). Also we do not split the data for \( \text{N2SID} \). We have also carried out experiments when we did split the data. This resulted in most cases in comparable results and in some cases even better results.

The SID methods are compared with the prediction error method \( \text{PEM} \) of the matlab System Identification toolbox Ljung [2007]. Here the involvement of the user in specifying the model structure is avoided by initializing \( \text{PEM} \) with the model determined by \( \text{N4SID} \). Therefore, the model order of \( \text{PEM} \) is the same as that determined by the \( \text{N4SID} \) method. In this way no user selection parameters are needed to be specified for \( \text{PEM} \). This is in agreement with the recommendation given on the \( \text{PEM} \) help page [http://nf.nci.org.au/facilities/software/Matlab/toolbox/ident/pem.html](http://nf.nci.org.au/facilities/software/Matlab/toolbox/ident/pem.html) when identifying black-box state space models.

### 6.3 Results and Discussion

The three SID methods in Table 3 and the \( \text{PEM} \) method will be compared for the data sets in Table 1. The results of this comparison are for each data set summarized in two graphs in the same figure. The left graph of the figure displays the goodness of fit criterium VAF. This is defined using the identified quadruple of system matrices \([\hat{A}_T, \hat{B}_T, \hat{C}_T, \hat{D}]\) obtained with each method to predict the output using the validation data set. Let the predicted output be denoted by \( \hat{y}_v(k) \) for each method, and let the output measurement in the validation data set be denoted by \( y_v(k) \). Then VAF

### Table 1. Five benchmark problems from the DaISy collection, De Moor et al. [1997]; \( N_{\text{tot}} \) is the total number of data samples available

| Nr | Data set     | Description                        | Inputs | Outputs | \( N_{\text{tot}} \) |
|----|-------------|------------------------------------|--------|---------|---------------------|
| 1  | 96-007      | CD player arm                      | 2      | 2       | 2048                |
| 2  | 98-002      | Continuous stirring tank reactor   | 1      | 2       | 7500                |
| 3  | 96-006      | Hair dryer                         | 1      | 1       | 1000                |
| 4  | 97-002      | Steam heat exchanger               | 1      | 1       | 4000                |
| 5  | 96-011      | Heat flow density                  | 2      | 1       | 1680                |

### Table 2. The increasing length \( N_{\text{ide}} \) of the data sets used for system identification starting with the sample index \( \text{del} \); \( N_{\text{val}} \) indicates the length of the validation data set starting with the sample index \( \max(N_{\text{ide}}) + 1 \).

| Nr | \( N_{\text{ide}} \) | \( \text{del} \) | \( N_{\text{val}} \) |
|----|----------------------|-----------------|---------------------|
| 1  | 80                   | 120             | 150                 |
| 2  | 100                  | 150             | 200                 |
| 3  | 80                   | 100             | 120                 |
| 4  | 150                  | 200             | 300                 |
| 5  | 175                  | 200             | 350                 |

\( \text{Nr} \) indicates the length of the validation data set starting with the sample index \( \max(N_{\text{ide}}) + 1 \). The value of \( N_{\text{val}} \) is listed in Table 2.
is defined as:

\[
\text{VAF} = \left(1 - \frac{1}{N_{\text{val}}} \sum_{k=1}^{N_{\text{val}}} \left\| y_{v}(k) - \hat{y}_{v}(k) \right\|^2 \right) 100\%
\]  

(26)

The right graph of the figure displays the model complexity as defined by the model order of the state space model. Both the goodness of fit and estimated model order are graphed versus the length of the identification data batch as indicated by the symbol \(N_{\text{ide}}\) in Table 2.

All these results are obtained in a similar “automized manner” for fair comparison as outlined in section 6.2. In order to evaluate the results additional information is retrieved from the singular values as computed by the SID methods \(\text{WNNopt}\) and \(\text{N2SID}\). This is done in order to see possible improvements in the low rank detection by the new SID method \(\text{N2SID}\) over \(\text{WNNopt}\). For an illustration of the potential improvement of the latter over \(\text{N4SID}\) we refer to Liu et al. [2013].

### The CD player arm data set (# 1 in Table 1)

The results are summarized in Figure 1. The goodness of fit is given on the left side of this figure and the detected order \(\hat{n}\) on the right side. For \(N_{\text{ide}} \leq 400\), \(\text{N2SID}\) outperforms all other methods and it was always better than \(\text{N4SID}\). \(\text{PEM}\) is able to improve the results of \(\text{N4SID}\) in most cases. Its results remain however inferior to \(\text{N2SID}\). In general \(\text{N2SID}\) detects a larger model order. For the shortest data lengths \(N_{\text{ide}} = 80\) and \(120\), \(\text{WNNopt}\) was not able to produce results since for that case the ADMM implementation broke down. The reason being that the Schur form was no longer computable as it contained NaN numbers. For that reason both the VAF and the order were put to zero. The \(\text{WNNopt}\) determined for \(150 \leq N_{\text{ide}} \leq 300\) a lower model order \(\hat{n}\) compared to \(\text{N2SID}\), but this at the cost of a lower VAF. For \(N_{\text{ide}} = 150\) and 175, the same order as for \(\text{N4SID}\) was detected, however with worsen VAF as compared to both \(\text{N4SID}\) and \(\text{PEM}\). For \(N_{\text{ide}} \geq 400\) the limit set on the model order, which was 10 in all experiments, was selected by \(\text{WNNopt}\), sometimes but not always yielding a better VAF.

The effect of the use of instrumental variables and the splitting of the identification data set to avoid overfitting on the order selection is clear from the singular values of \(\text{N2SID}\) and \(\text{WNNopt}\) given in Figure 2 for \(N_{\text{ide}} = 600\). This plot visually supports the selection of a 7-th order model by \(\text{N2SID}\) and it also explains why \(\text{WNNopt}\) selects a larger model order. One possible explanation is that the instrumental variables and projections have “projected away” crucial information in the data.

### The Continuous stirred Tank Reactor data set (# 2 in Table 1)

The goodness of fit parameter VAF and the estimated model order \(\hat{n}\) are plotted in Figure 3 in the left and right graphs, respectively. For \(N_{\text{ide}} = 100\) and 150 \(\text{WNNopt}\) was not able to provide numerical results. For that reason the corresponding VAF values are again fixed to zero. \(\text{PEM}\) resulted in bad VAF results for \(N_{\text{ide}} = 100\), probably as a consequence of bad initialization from \(\text{N4SID}\). However, also for \(N_{\text{ide}} = 800\) \(\text{PEM}\) had severely degraded results despite the fact that \(\text{N4SID}\) provided a model of comparable quality with the other SID methods.

The singular values in Figure 4 indicate that for \(N_{\text{ide}} = 800\) both \(\text{N2SID}\) and \(\text{WNNopt}\) have the same order estimate \(\hat{n}\). There is a clear gap in the singular values for \(\text{WNNopt}\). The difference in detected order despite similar VAF indicates that order detection is not so critical for this example.
The Hair dryer data set (\# 3 in Table 1)  
The goodness of fit parameter VAF and the estimated model order $\hat{n}$ are plotted in Figure 7.  
$N2SID$ again for small data sets with $N_{ide}$ ranging between 150 and 750 yields the best output predictions of all methods. Comparing the VAF value in Figure 7 with those for the previous Daisy data sets reveals that the values are smaller. This reflects problems with the data set due to lower signal to noise ratio, system nonlinearity, etc. Because of this we started the analysis with the smallest value of $N_{ide}$ equal to 150, since for smaller values poor results were obtained for all methods.  
The other methods show a similar behavior as for the previously analysed data sets: in most cases but not all PEM improves over $N4SID$, $WNNopt$ displays inferior behavior for $N_{ide} \leq 1250$, and both $N2SID$ and $N4SID$ (PEM) determine a smaller order then $WNNopt$.

Figure 6 displays the singular values for the last data set where $N_{ide} = 400$ in Figure 6. It confirms the improved potential in low rank approximation by $N2SID$ over $WNNopt$. The latter method diminishes the gap, leading in general to a larger model order estimation. This larger model order does however for this example not lead to a better output prediction.

The Steam Heat Exchanger data set (\# 4 in Table 1)  
The goodness of fit parameter VAF and the estimated model order $\hat{n}$ are plotted in Figure 7.  
$N2SID$ again for small data sets with $N_{ide}$ ranging between 150 and 750 yields the best output predictions of all methods. Comparing the VAF value in Figure 7 with those for the previous Daisy data sets reveals that the values are smaller. This reflects problems with the data set due to lower signal to noise ratio, system nonlinearity, etc. Because of this we started the analysis with the smallest value of $N_{ide}$ equal to 150, since for smaller values poor results were obtained for all methods.

The other methods show a similar behavior as for the previously analysed data sets: in most cases but not all PEM improves over $N4SID$, $WNNopt$ displays inferior behavior for $N_{ide} \leq 1250$, and both $N2SID$ and $N4SID$ (PEM) determine a smaller order than $WNNopt$.  

Figure 6 displays the singular values for the last data set where $N_{ide} = 400$ in Figure 6. It confirms the improved potential in low rank approximation by $N2SID$ over $WNNopt$. The latter method diminishes the gap, leading in general to a larger model order estimation. This larger model order does however for this example not lead to a better output prediction.
Finally, the plot of the singular values for the last data set in Figure 8 displays a similar behavior. Both singular value plots clearly support the automatic order selection made. However, N2SID has a better trade-off between model complexity and model accuracy as expressed by the VAF.

![Steam heat exchanger](image1)

Fig. 7. VAF Daisy # 4 - Steam Heat Exchanger.

Heat flow density data set (# 5 in Table 1) The goodness of fit parameter VAF and the estimated model order \( \hat{n} \) are plotted in Figure 9.

For this data set, \textit{WNNopt} provides for \( 200 \leq N_{ide} \leq 450 \) the best results but in general detects a larger model order. For the smallest length data set \textit{WNNopt} produced inferior VAF. N2SID provides a better VAF prediction compared to N4SID and PEM and this for a smaller model order \( \hat{n} \) as compared to \textit{WNNopt}.

From the singular values in Figure 10 for \( N_{ide} = 600 \) an order selection of 1 up to 4 is clearly justified by N2SID. The order selection made by \textit{WNNopt} is much less clear.

General Observation from the analysed Daisy data sets.

The automated analysis of the 5 Daisy data sets clearly demonstrates the merit of the new SID method N2SID over the other representative identification methods considered. Especially when considering data sets of small length it is able to make a good and sometimes excellent trade-off between model complexity and model accuracy as expressed by the goodness of fit. The improvement over the other analysed nuclear norm subspace identification method \textit{WNNopt} in order detection both in revealing a clear gap as well as in detecting models of low complexity is evident.

7. CONCLUDING REMARKS

Subspace identification is revisited in this paper in the scope of nuclear norm optimization methods. A new way to impose structural matrix properties in the data equation in subspace identification on the measured data has been presented. The new subspace identification method is referred to as N2SID. It is shown that, especially for small data length batches, when the number of samples is only a small multiple of the order of the underlying system, the incorporation of structural information about the low rank
property of the matrix revealing the required subspace, and about the block Toeplitz structure of the matrix containing the unknown Markov parameters enables to improve the results of widely used SID methods and of the prediction error method. In addition to the structural constraints the N2SID method also enables to make a trade-off in the first step of the calculations between the subspace approximation and the prediction error cost function. Because of this it overcomes the drawback that SID did not consider a classical prediction error cost function.

We have also shown how to implement efficient code for N2SID by making use of the ADMM algorithm for solving nuclear norm optimization problems. This makes it possible to use N2SID not only for short data batches. The single integrative step that aims at imposing key structural matrix properties makes a trade-off between the prediction error cost function and the problem to retrieve the subspace of interest. This integrated approach may help to simplify the analysis of the optimality of SID methods and to further clarify their link with prediction error methods. This new way of looking upon SID will open up the possibility for new developments in the future. One such development is that it may help in further automizing the way linear models are retrieved from experimental data sets. We also mention the trivial extension to missing output data case.

Acknowledgement: The authors kindly acknowledge Mr. Baptiste Sinquin from Ecole Centrale Lyon for his help in a preliminary matlab comparison study with N2SID during his internship at the Delft Center for Systems and Control under the supervision of Prof. M. Verhaegen.

REFERENCES

A. Chiuso. The role of vector autoregressive modeling in predictor-based subspace identification. Automatica, 43 (6):1034 – 1048, 2007.

B. De Moor, P. De Gersem, B. De Schutter, and W. Favoreel. DAISY: A database for the identification of systems. Journal A, 38(3):4–5, Sep. 1997.

M. Fazel. Matrix Rank Minimization with Applications. PhD thesis, Stanford University, 2002.

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, pages 4734–4739, 2001.

M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. 2012. Submitted for possible publication.

G. H. Golub and Ch. F. van Loan. Matrix Computations. The Johns Hopkins University Press, Baltimore, Maryland, 1996. Third Edition.

A. Hansson, Z. Liu, and L. Vandenberghe. Subspace system identification via weighted nuclear norm optimization. In Proceedings of the 51st IEEE Conference on Decision and Control, 2012. Submitted.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications, 31(3):1235–1256, 2009b.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications, 31(3):1235–1256, 2010.

Z. Liu, A. Hansson, and L. Vandenberghe. Nuclear norm system identification with missing inputs and outputs. Systems & Control Letters, 62:605–612, 2013.

L. Ljung. System Identification — Theory for the User. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

L. Ljung. System Identification Toolbox for use with MATLAB. Version 7. The MathWorks, Inc, Natick, MA, 7th edition, 2007.

K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In Proceedings of American Control Conference, pages 2953–2959, 2010.

T. Roh and L. Vandenberghe. Discrete transforms, semidefinite programming, sum-of-squares representations of nonnegative polynomials. SIAM Journal on Optimization, 16(4):939–964, 2006.

R.S. Smith. Nuclear norm minimization methods for frequency domain subspace identification. Proceedings of the American Control Conference, pages 2689–2694, 2012.

H. Hjalmarsson and J. S. Welsh and C. R. Rojas. Identification of box-jevkin models using structured arx models and nuclear norm relaxation. Proceedings of the 16th IFAC Symposium on System Identification, pages 322–327, 2012.

B. Wahlberg and C.R. Rojas. Matrix Rank Optimization Problems in System Identification via Nuclear Norm Minimization. (Tutorial) at the European Control Conference, 2013.

P. van Overschee and B. De Moor. Subspace Identification of Linear Systems: Theory, Implementation, Applications. Springer Publishing, 1996.

Z. Liu, A. Hansson, and L. Vandenberghe. Semidefinite programming methods for system realization and identification. In Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, pages 4676–4681, 2009a.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications, 31(3):1235–1256, 2009b.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications, 31(3):1235–1256, 2010.

Z. Liu, A. Hansson, and L. Vandenberghe. Nuclear norm system identification with missing inputs and outputs. Systems & Control Letters, 62:605–612, 2013.

L. Ljung. System Identification — Theory for the User. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

L. Ljung. System Identification Toolbox for use with MATLAB. Version 7. The MathWorks, Inc, Natick, MA, 7th edition, 2007.

K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In Proceedings of American Control Conference, pages 2953–2959, 2010.

T. Roh and L. Vandenberghe. Discrete transforms, semidefinite programming, sum-of-squares representations of nonnegative polynomials. SIAM Journal on Optimization, 16(4):939–964, 2006.

R.S. Smith. Nuclear norm minimization methods for frequency domain subspace identification. Proceedings of the American Control Conference, pages 2689–2694, 2012.

H. Hjalmarsson and J. S. Welsh and C. R. Rojas. Identification of box-jevkin models using structured arx models and nuclear norm relaxation. Proceedings of the 16th IFAC Symposium on System Identification, pages 322–327, 2012.

B. Wahlberg and C.R. Rojas. Matrix Rank Optimization Problems in System Identification via Nuclear Norm Minimization. (Tutorial) at the European Control Conference, 2013.

P. van Overschee and B. De Moor. Subspace Identification of Linear Systems: Theory, Implementation, Applications. Springer Publishing, 1996.

Z. Liu, A. Hansson, and L. Vandenberghe. Semidefinite programming methods for system realization and identification. In Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, pages 4676–4681, 2009a.