Available online at www.sciencedirect.com
ScienceDirect
Journal of Process Control 18 (2008) 582–593
JOURNAL OF PROCESS CONTROL
www.elsevier.com/locate/jprocont

Comparison of input signals in subspace identification of multivariable ill-conditioned systems

Andrea Micchi, Gabriele Pannocchia *

Department of Chemical Engineering, Industrial Chemistry and Science of Materials, University of Pisa, Via Diotisalvi, 2, 56126 Pisa, Italy

Received 6 November 2006; received in revised form 3 October 2007; accepted 9 October 2007

Abstract

Ill-conditioned processes often produce data of low quality for model identification in general, and for subspace identification in particular, because data vectors of different outputs are typically close to collinearity, being aligned in the "strong" direction. One of the solutions suggested in the literature is the use of appropriate input signals, usually called "rotated" inputs, which must excite sufficiently the process in the "weak" direction. In this paper open-loop (uncorrelated and rotated) random signals are compared against inputs generated in closed-loop operation, with the aim of finding the most appropriate ones to be used in multivariable subspace identification of ill-conditioned processes. Two multivariable ill-conditioned processes are investigated and as a result it is found that closed-loop identification gives superior models, both in the sense of lower error in the frequency response and in terms of higher performance when used to build a model predictive control system.

Keywords: Input design; Subspace identification; Ill-conditioned systems; Model predictive control

1. Introduction and previous work

During the last decades, a large number of valuable contributions were brought in system identification, especially for the definition of consistent, reliable and numerically efficient identification algorithms. In general, system identification can be performed using two different approaches: Prediction Error methods (PE) and Subspace IDentification methods (SID). The first approach (PE) is based on the simple idea that, in absence of disturbances, the better the model recovered from data, the smaller the prediction error it generates, where the prediction error is the difference between the measured output and the model output data. The second approach (SID) instead involves particular matrices obtained from output and input data and performs projection operations to cancel out the noise contributions. Thus, the system model is obtained in state-space form using these projected data matrices. PE methods were principally developed and probably called with this name for the first time by Ljung (see [1] for a complete overview of PE algorithms), but, even before his works, it was common to minimize the prediction error for identification of process parameters. SID is a relatively young technique, developed mainly during the last fifteen years, even if its basis were posed several years before. Van Overschee and De Moor [2] and Verhaegen [3] introduced and solved the subspace problem using different approaches, called N4SID and MOESP, respectively, even though it was shown later in [4] that those methods use the same subspace, and that they only differ in the weighting matrices. Another successful class of subspace methods is that referred to as CVA algorithms, developed by Larimore [5]. During the years, several modifications and improvements were made on those algorithms with the aim of enhancing numerical stability and efficiency. In particular, several new subspace approaches were recently developed,

* Corresponding author. Tel.: +39 050 511238; fax: +39 050 511266.
E-mail address: g.pannocchia@ing.unipi.it (G. Pannocchia).

0959-1524/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2007.10.003

such as those from Wang and Qin [6] and Huang et al. [7]. This latter algorithm is that used in the present work, with some modifications introduced to improve results in the computation of the system matrices.

As remarked by Zhu [8], however, a fundamental portion of the "identification problem", that is test design, received less attention. This issue was studied for the single-input single-output (SISO) case by Gevers and Ljung [9], who showed that it was possible/desirable to design inputs in dependence of intended model application. After this work, Koung and MacGregor [10,11] addressed the problem of optimal test design and system identification for robust control of multi-input multi-output (MIMO) processes. Hjalmarsson et al. [12] emphasized that models of superior quality are obtained from closed-loop tests, particularly for model-based control purposes. Nonetheless, during these years (and in a large number of cases still at present days) standard industrial inputs for system identification were considered to be step tests. This approach is, in general, suboptimal because output data obtained in such a way lack of information about the system to be identified, since the inputs are rather poor in terms of frequency content. Alternatively, one can use binary random or pseudo-random inputs, because of their superior power spectrum with respect to that of step signals [1,13].

Unlike SISO processes, MIMO systems may show "directions" (in the input vector space) in which the (steady-state or dynamic) effect of the inputs on the process outputs is much larger than in other directions. In such situations the process is said to be ill-conditioned, and frequent examples of ill-conditioned systems are high-purity distillation columns. Ill-conditioned processes are usually difficult to be controlled, because decentralized controllers are typically inadequate due to large interactions among the control loops, but multivariable model-based controllers may suffer from robustness issues [14,15]. Moreover, ill-conditioned processes may be difficult to deal with from an identification point of view, because traditional uncorrelated open-loop step tests tend to excite the system mostly in high-gain directions [10,16]. In fact, depending on the input directions, the output response can vary significantly in magnitude, even thousands times from the high-gain to the low-gain direction. Thus, the information coming out from the high-gain direction is predominant over that from the low-gain direction, often resulting in a model not suitable for control purposes [11,17].

Koung and MacGregor [10,11] developed a test design method using highly correlated input signals, which are able to excite the process both in high-gain and low-gain directions. However, highly correlated input signals may constitute a major trouble when SID methods are used, especially for system order recovering: in fact, matrices used for identification could be near to rank deficiency in systems of this type, and so algorithms could lead to incorrect numerical results [18]. Misra and Nikolaou [19] also presented a work dealing with this issue: differently from [10,11], they focused only on order determination (and not on model recovering), and introduced the problem from the point of view of subspace identification. In particular, they proposed an open-loop test design based on "rotated" inputs, in which the angles between the inputs are obtained by trial and error. Conner and Seborg [20], similarly, indicated the same rotated inputs as the best solution for system identification. Other relevant contributions on test design were given by Stec and Zhu [21] who introduced a quite different design, composed by a part of high magnitude collinear inputs, used to estimate characteristics of the low-gain direction, and small uncorrelated inputs, ideal to estimate properties of the process in high-gain direction, by Cooley and Lee [22] who obtained control-relevant finite-impulse response (FIR) models, by Gopaluni et al. [23] who proposed a test design particularly suitable to recover good Model Predictive Control (MPC) oriented models (minimizing j-step ahead prediction error). Stec and Zhu expanded their work in [24], where they introduced another kind of test design, in which the input signal is the sum of a high amplitude linearly dependent signal and a low amplitude uncorrelated signal. Hjalmarsson [25] presented an all-inclusive paper spanning the system identification problem, test design included. Recently, Bruwer and MacGregor [26] extended the work in [10,11] to the (frequent) case of presence of physical and process constraints, also considered in a recent work by Zhan et al. [27].

The main objective of this work is the comparison of different test signals for subspace identification of ill-conditioned systems. In particular, advantages and disadvantages of rotated inputs are analyzed and these characteristics are compared with those of closed-loop signals. The identified models are then compared in terms of frequency error analysis and closed-loop performance when used in a model predictive control algorithm. The rest of this paper is organized as follows. In Section 2 the subspace identification method used in the present work is described in details. Section 3 is focused on the analysis of the different types of signals with particular attention to their effects on system order recovering and overall quality of the identified models. Moreover, a generalization of the rotated inputs design procedure to non-square multivariable systems of arbitrary dimensions is presented. In Section 4 two case studies of ill-conditioned distillation columns are reported to show the results obtained using different kinds of signals and discuss the problems that could be experienced using rotated inputs. Finally, Section 5 summarizes the main achievements of this work, namely that closed-loop inputs give better models than open-loop rotated inputs.

2. Subspace identification method

2.1. Basic definitions

In this paper, linear discrete-time-invariant state-space systems in the following form are considered:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Ke_k \\ y_k &= Cx_k + e_k, \end{aligned} \tag{1}$$

in which $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, $y \in \mathbb{R}^p$ is the output, $e \in \mathbb{R}^p$ is stochastic noise, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ are the system matrices and $K \in \mathbb{R}^{n\times p}$ is the noise model matrix. Throughout this paper the superscript $'$ represents the transpose operator. The following standing assumptions are considered.

Assumption 1. The pair (A, B) and (A, K) are stabilizable, the pair (A, C) is observable, the closed-loop matrix $(A - KC)$ is strictly Hurwitz (in a discrete-time sense), the noise e_k is white, and statistically independent of past outputs and inputs, i.e. $E\{y_k e_j'\} = 0$ and $E\{u_k e_j'\} = 0$ for all $j > k$.

Given a positive integer r, assumed to satisfy $r > n$, let the vectors of "future" (with respect to time k) outputs, inputs and noises be defined, respectively, as:

$$\boldsymbol{y}_k = \begin{bmatrix} y_k \\ y_{k+1} \\ \vdots \\ y_{k+r-1} \end{bmatrix}, \quad \boldsymbol{u}_k = \begin{bmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+r-1} \end{bmatrix}, \quad \boldsymbol{e}_k = \begin{bmatrix} e_k \\ e_{k+1} \\ \vdots \\ e_{k+r-1} \end{bmatrix}. \tag{2}$$

Assumption 2. Data vectors (u, y) are collected for L sampling times, namely from sample time 0 to sample time $L - 1$ (with $L \gg r$).

From the model (1), one can obtain:

$$\boldsymbol{y}_k = \Gamma_r x_k + H_r^u \boldsymbol{u}_k + H_r^e \boldsymbol{e}_k, \tag{3}$$

in which Γ_r is the extended observability matrix, H_r^u and H_r^e are lower block-triangular Toeplitz matrices:

$$\Gamma_r = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{r-1} \end{bmatrix}, \quad H_r^u = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ CB & 0 & \cdots & \cdots & 0 \\ CAB & CB & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ CA^{r-2}B & \cdots & \cdots & CB & 0 \end{bmatrix},$$

$$H_r^e = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ CK & I & \cdots & \cdots & 0 \\ CAK & CK & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ CA^{r-2}K & \cdots & \cdots & CK & I \end{bmatrix}, \tag{4}$$

in which 0 and I are used to denote the full zero matrix and the identity matrix, respectively, of suitable dimensions.

Being the data collected for L sampling times, it is possible to write:

$$Y_f = \Gamma_r X + H_r^u U_f + H_r^e E_f, \tag{5}$$

where the matrices Y_f, X, U_f, E_f are constructed by placing side-by-side the vectors $\boldsymbol{y}_k, x_k, \boldsymbol{u}_k$ and $\boldsymbol{e}_k$ respectively, i.e.

$$\begin{aligned} Y_f &= [\boldsymbol{y}_r \quad \boldsymbol{y}_{r+1} \quad \cdots \quad \boldsymbol{y}_{r+M-1}], \quad X = [x_r \quad x_{r+1} \quad \cdots \quad x_{r+M-1}], \\ U_f &= [\boldsymbol{u}_r \quad \boldsymbol{u}_{r+1} \quad \cdots \quad \boldsymbol{u}_{r+M-1}], \quad E_f = [\boldsymbol{e}_r \quad \boldsymbol{e}_{r+1} \quad \cdots \quad \boldsymbol{e}_{r+M-1}], \end{aligned} \tag{6}$$

in which the number of columns of these matrices is equal to $M = L - 2r + 1$. In some subspace identification algorithms, as the one considered in this work, it is common to introduce the matrix $Z_f \in \mathbb{R}^{(pr+mr)\times M}$, which is obtained by stacking Y_f and U_f, that is

$$Z_f = \begin{bmatrix} Y_f \\ U_f \end{bmatrix}. \tag{7}$$

Similarly, the matrix $Z_p = [\, Y_p' \quad U_p' \,]'$ can be defined using the matrices of "past" output and input data, i.e. $Y_p = [\boldsymbol{y}_0 \quad \boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_{M-1}]$ and $U_p = [\boldsymbol{u}_0 \quad \boldsymbol{u}_1 \quad \cdots \quad \boldsymbol{u}_{M-1}]$.

2.2. System matrices recovering

Subspace methods perform a projection of output (and input) data onto a subspace orthogonal to noise: in this way, the projected data matrices depend only on deterministic contributions, while all the stochastic information contained in the original data is discarded. Each method differs from the others in the matrices used to perform these projections and in the way the model matrices are obtained from Γ_r, H_r^u and H_r^e.

In this work, the method of orthogonal projections proposed by Huang et al. [7] is used, introducing some modifications as detailed later in this section. The adopted algorithm is particularly interesting, because it overcomes the main problem of many subspace methods, that is the lack of consistency when used with closed-loop (CL) data. This feature is essential in the present work, because data collected in closed-loop are shown to be very effective in subspace identification of ill-conditioned systems.

The underlying philosophy of the method is to recover the matrix Γ_r from the (left) orthogonal space of the matrix Z, defined as:

$$Z = Z_f W, \tag{8}$$

where two choices of the orthogonal projection matrix $W = W' \in \mathbb{R}^{M\times M}$ are considered (see [7] for more details on orthogonal projections), depending on the data collection scheme. If data are collected in open loop, the appropriate choice is $W = Z_p^+ Z_p$, where the superscript $^+$ denotes the right pseudo-inverse operator (computed via singular value decomposition, SVD). Huang et al. [7] show that this choice leads to an orthogonal projection of Z_f onto the row space of Z_p. If data are collected in closed-loop, instead, the appropriate choice is $W = Z_{CL}^+ Z_{CL}$, where $Z_{CL} = [\, \overline{Y}_f' \quad Z_p' \,]'$ and $\overline{Y}_f \in \mathbb{R}^{pr\times M}$ is the matrix of future setpoints, defined similarly to Y_f [7].

Having defined the projected data matrix Z, the first step is to perform an SVD:

$$Z = U_Z S_Z V_Z' = [\, U_1 \quad U_2 \,] \begin{bmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \end{bmatrix} V_Z', \tag{9}$$

where S_1 and S_2 are diagonal matrices, which contain the significant and the negligible singular values of Z, respectively. The dimension of S_1 can obtained by using e.g. an Akaike information criterion as in [6] or a heuristic principal component analysis (PCA) approach as described. The first mr singular values of Z are considered, and then subsequent $\hat{n}$ significant singular values are selected according to:

$$\frac{\sigma_{mr+\hat{n}}}{\sum_{j=1}^{\hat{n}}\sigma_{mr+j}} > \rho, \quad \sigma_j \in \text{diag } S_Z, \quad 1 \leqslant \hat{n} \leqslant r, \tag{10}$$

in which ρ is a positive scalar close to 0 (typical values for ρ are between 0.01 and 0.05). Thus, the diagonal matrix S_1, containing the significant singular values, has dimension $mr + \tilde{n}$, where $\tilde{n}$ is considered equal to the largest value of $\hat{n}$ for which (10) holds. Notice that since the left singular vectors matrix of Z, i.e. U_Z, has dimension $mr + pr$, the matrix U_2 in (9) has $pr - \tilde{n}$ columns. Huang et al. [7] show that:

$$\Gamma_r^{\perp}[I \quad -H_r^u] = TU_2', \tag{11}$$

where $\Gamma_r^{\perp}$ is a basis matrix for the left null space of Γ_r, i.e. a full rank matrix such that $(\Gamma_r^{\perp})\Gamma_r = 0$ and T is any non-singular transformation matrix of suitable dimensions (often chosen as the identity matrix). Eq. (11) and a suitable partitioning:

$$TU_2' = [P_1' \quad P_2'], \tag{12}$$

in which P_1' has pr columns, allow one to compute Γ_r and H_r^u from the following relations:

$$P_1'\Gamma_r = 0 \tag{13a}$$

$$-P_1'H_r^u = P_2'. \tag{13b}$$

It is clear that Γ_r is computed from (13a) as an orthonormal basis matrix of the (right) null space of P_1', whereas (13b) is solved for H_r^u in a least-squares sense. Notice that the computation of H_r^u in a least-squares sense can be improved by taking into account the Toeplitz structure of this matrix, reported in (4).

The matrices A and C are computed from Γ_r by using the shift invariance property of the observability matrix, that is (using a Matlab notation):

$$C = \Gamma_r(1:p,:) \tag{14a}$$

$$\Gamma_r(1:(r-1)p,:)A = \Gamma_r(p+1:rp,:), \tag{14b}$$

where (14b) is solved for A in a least-squares sense. In this work, computation of the matrix B is changed with respect to the original algorithm, because several studies pointed out poor results of this step in a number of cases [28,29]. In particular, Huang et al. [7, Sec. 3.1] recovered the matrix B from H_r^u, referring to [6] for this step. A different approach to recover B was proposed by Qin et al. [28] and independently by Pannocchia et al. [29], which is very similar to that used in [1]. Specifically, it consists in obtaining that matrix by using a prediction error approach and solving a least-squares problem. The details are omitted in the sake of space, but the interested reader is referred to the cited papers [28,29].

Finally, if required, the disturbance model matrix K can be obtained in several ways. Huang et al. [7, Sec. 3.3] propose a formulation based on Kalman filter states computed from Γ_r and H_r^u, which is applicable to open-loop data only. They also discuss an alternative formulation applicable to closed-loop data, which in turn computes B, C and K in a slightly different way.

2.3. Open-loop and closed-loop test design

Open-loop (OL) tests are widely used in industrial system identification, even if CL tests may be preferred for some practical and theoretical reasons. At first CL tests, unlike OL tests, can be used in case the outputs have to be maintained within a range, so it is possible to perform tests for data collection while maintaining products at the desired specifications. Moreover, CL test data usually contain information over the most significant frequencies for control, so a model obtained with this kind of data is often superior when used for model-based control, as reported in [9]. In the end, CL test data seem to be superior in system model recovery, especially for ill-conditioned process as it is discussed later in Section 4.

However, some problems could be experienced with CL test data when subspace identification is performed, but they can be overcome with specific modifications of the algorithms (see [30] for example). In particular the orthogonal projection method used in this work can be applied successfully to closed-loop data provided that the orthogonal projection matrix W in (8) is appropriately chosen, as previously outlined in Section 2.2. The interested reader is referred to [7, Sec. 3.2] for further details on closed-loop data treatment.

3. Design and analysis of input signals for ill-conditioned processes

Ill-conditioned systems, as previously said, represent one of the most difficult kind of linear processes to be identified: *gains* appear very different depending on input direction, so several difficulties could be experienced. Models obtained for these processes often suffer from significant errors in order and gain recovery. Indeed, it can be seen in [19] that matrices used for subspace identification in a lot of cases have nearly collinear rows, and this can lead to incorrect order estimation and subsequently in erroneous models.

3.1. Design of rotated inputs

In order to avoid the generation of output data mostly aligned in the strong direction, Koung and MacGregor [11] and Misra and Nikolaou [19] propose the use of "rotated" inputs, in which the input vectors are strongly aligned in pre-specified directions. For a 2×2 system, this

means that the input vectors form particular angles w.r.t. the principal axes in a (u_1, u_2) plane, and these angles are referred to as "rotation angles". The algorithm for the generation of rotated inputs in the 2×2 case [11,19] is here extended to the most general case of non-square multivariable systems of arbitrary dimensions.

Let $G = C(I-A)^{-1}B \in \mathbb{R}^{p\times m}$ be the gain matrix of the system, and consider the steady-state relation $y^s = G\ u^s$, in which from now on the superscript ^{s} indicates *steady state*, which can be written as follows by means of an SVD:

$$y^s = U_G S_G V_G' u^s = \begin{bmatrix} \mathbf{u}_{11} & \cdots & \mathbf{u}_{1t} \\ \vdots & & \vdots \\ \mathbf{u}_{p1} & \cdots & \mathbf{u}_{pt} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{s}_2 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{s}_t \end{bmatrix} \begin{bmatrix} \mathbf{v}_{11} & \cdots & \mathbf{v}_{m1} \\ \vdots & & \vdots \\ \mathbf{v}_{1t} & \cdots & \mathbf{v}_{mt} \end{bmatrix} \begin{bmatrix} u_1^s \\ \vdots \\ u_m^s \end{bmatrix} \tag{15a}$$

$$= \begin{bmatrix} \mathbf{u}_{11} & \cdots & \mathbf{u}_{1t} \\ \vdots & & \vdots \\ \mathbf{u}_{p1} & \cdots & \mathbf{u}_{pt} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1(\mathbf{v}_{11}u_1^s + \cdots + \mathbf{v}_{m1}u_m^s) \\ \vdots \\ \mathbf{s}_t(\mathbf{v}_{1t}u_1^s + \cdots + \mathbf{v}_{mt}u_m^s) \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{11} & \cdots & \mathbf{u}_{1t} \\ \vdots & & \vdots \\ \mathbf{u}_{p1} & \cdots & \mathbf{u}_{pt} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_t \end{bmatrix} \tag{15b}$$

$$= \xi_1 \begin{bmatrix} \mathbf{u}_{11} \\ \vdots \\ \mathbf{u}_{p1} \end{bmatrix} + \cdots + \xi_t \begin{bmatrix} \mathbf{u}_{1t} \\ \vdots \\ \mathbf{u}_{pt} \end{bmatrix}, \tag{15c}$$

in which $t = \min(p,m)$, i.e. the minimum between the number of outputs and the number of inputs of the system. For ill-conditioned systems, since $\mathbf{s}_1 \gg \mathbf{s}_t$ if a sequence input vectors u^s is not *carefully* generated (e.g. a random or pseudo-random input vector sequence), it is easy to see that the sequence of output vectors (at steady-state) is mostly aligned over the space spanned by the first columns of the left singular vectors matrix U_G, i.e. in the stronger directions. This occurs because, unless special care is taken in choosing the input vectors, it follows that $|\xi_1| \gg |\ \xi_t|$ and hence the last terms in the sum (15c) become negligible, so that no information about the system's weaker directions is contained in the data. In the spirit of [11], here the goal is to construct a vector (or a sequence of vectors) u^s such that the corresponding (steady-state) output vector contains information regarding all singular values in *equal* magnitude. Since the columns of the left singular vector matrix U_G are normalized, such a goal can be achieved by imposing the following $t-1$ conditions[1]:

$$\xi_1 = \xi_2 = \ldots = \xi_t, \tag{16}$$

which can be explicitly written as:

$$\underbrace{\begin{bmatrix} v_{1,2} & \cdots & v_{m,2} \\ \vdots & & \vdots \\ v_{1,t} & \cdots & v_{m,t} \end{bmatrix}}_{\Theta} \begin{bmatrix} u_1^s \\ \vdots \\ u_m^s \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{17}$$

where $v_{\alpha,\beta} = s_1\mathbf{v}_{\alpha,1} - s_\beta \mathbf{v}_{\alpha,\beta}$.

The linear problem (17) consists of $t-1$ equations in m variables; so let $z = m - t + 1$ be the number of degrees of freedom, i.e. the number of input components that can be chosen arbitrarily. From the definition of t, it is straightforward to see that $z \geqslant 1$. Assume that the first z components of u^s are chosen arbitrarily, the remaining $m - z = t - 1$ components can be computed so that (17) is satisfied by solving the following square linear system:

$$\underbrace{\begin{bmatrix} -v_{z+1,2} & \cdots & -v_{m,2} \\ \vdots & & \vdots \\ -v_{z+1,t} & \cdots & -v_{m,t} \end{bmatrix}}_{\Phi} \begin{bmatrix} u_{z+1}^s \\ \vdots \\ u_m^s \end{bmatrix} = \underbrace{\begin{bmatrix} v_{1,2} & \cdots & v_{z,2} \\ \vdots & & \vdots \\ v_{1,t} & \cdots & v_{z,t} \end{bmatrix}}_{\Psi} \begin{bmatrix} u_1^s \\ \vdots \\ u_z^s \end{bmatrix}, \tag{18}$$

where the vector $[\,u_{z+1}^s \quad \ldots \quad u_m^s\,]'$ contains the unknowns. Assuming that Φ is invertible, the vector of unknown input components can be computed as:

$$\begin{bmatrix} u_{z+1}^s \\ \vdots \\ u_m^s \end{bmatrix} = \begin{bmatrix} -v_{z+1,2} & \cdots & -v_{m,2} \\ \vdots & & \vdots \\ -v_{z+1,t} & \cdots & -v_{m,t} \end{bmatrix}^{-1} \begin{bmatrix} v_{1,2} & \cdots & v_{z,2} \\ \vdots & & \vdots \\ v_{1,t} & \cdots & v_{z,t} \end{bmatrix} \begin{bmatrix} u_1^s \\ \vdots \\ u_z^s \end{bmatrix}$$

$$= \Phi^{-1}\Psi \begin{bmatrix} u_1^s \\ \vdots \\ u_z^s \end{bmatrix}. \tag{19}$$

In conclusion, a strategy for the design of a rotated input sequence is the following, which represents an extension

[1] Notice that (16) is only one of the possible 2^{t-1} valid combinations of the most general conditions: $\xi_1 = \pm\,\xi_2 = \ldots = \pm\,\xi_t$.

of the procedure proposed in [19] to higher dimension (possibly non-square) systems.

1. Define a random (or pseudo-random) multivariable binary signal $[u_{1,k} \quad \dots \quad u_{z,k}]'$ of dimension $z = m - t + 1$, for $k = 0, \dots, L-1$.
2. Perform an SVD of the system gain matrix G (or an approximation to it), compute the matrices Φ^{-1} and Ψ defined in (18).
3. For each sample time k, compute the remaining $t-1$ input components as:

$$\begin{bmatrix} u_{z+1,k} \\ \vdots \\ u_{m,k} \end{bmatrix} = \Phi^{-1}\Psi \begin{bmatrix} u_{1,k} \\ \vdots \\ u_{z,k} \end{bmatrix} + \zeta_k, \tag{20}$$

in which $\zeta \in \mathbb{R}^{t-1}$ is a white noise signal with "small" amplitude (dithering), used to avoid exact collinearity of the inputs.

Notice that, in case Φ is not invertible, a possible alternative is to compute a full rank matrix $\mathcal{N}_\Theta \in \mathbb{R}^{m \times z}$ such that $\Theta \mathcal{N}_\Theta = 0$ and to define, for each sampling time k, the whole input vector as:

$$\begin{bmatrix} u_{1,k} \\ \vdots \\ u_{m,k} \end{bmatrix} = \mathcal{N}_\Theta \mu_k + \zeta_k,$$

in which $\mu \in \mathbb{R}^z$ is the primary random (or pseudo-random) multivariable binary signal and $\zeta \in \mathbb{R}^m$ is a white noise (dithering) signal.

3.2. Issues in model order recovery

As discussed in Section 2, the considered subspace algorithm performs linear projections of a matrix Z_f formed by output and input data. In absence of noise, by its own construction (see Eqs. (3) and (6)), it immediately follows that:

$$\begin{aligned} \text{rank } Z_f &= \text{rank}\left(\begin{bmatrix} Y_f \\ U_f \end{bmatrix}\right) \\ &= \text{rank}\left(\begin{bmatrix} \Gamma_r & H_r^u \\ 0 & I \end{bmatrix}\begin{bmatrix} X_f \\ U_f \end{bmatrix}\right) \leqslant mr + n, \end{aligned}$$

in which the equality sign holds under appropriate conditions on the input excitation [6, Lemma 1]. An SVD performed on[2] Z_f leads to $mr + \tilde{n}$ non-zero singular values, where $\tilde{n} \leqslant n$. As discussed in Section 3.1 and also remarked in Section 4.2, for ill-conditioned plants when open-loop random inputs are used, the sequences of outputs are close to collinearity. Therefore, Z_f shows a number of significant singular values less than $m\,r + n$, the selected model order $\tilde{n}$ is less than n, and the model will be missing information about the weaker directions. In the presence of noise, the number of significant singular values (and therefore the selected model order), may increase but still no information regarding the weaker directions is obtained, because the random open-loop sequence of inputs did not excite the plant sufficiently in those directions.

As discussed, one of the solutions (and probably the most popular in the academic literature) that were proposed in the last years to avoid problems of order recovering in ill-conditioned processes is the use of "rotated" inputs (see [11,19]). Such input design was generalized to (possibly non-square) systems of arbitrary dimension in Section 3.1. However, the rank analysis previously described emphasizes that if the input vectors are strongly aligned in particular directions, as required by the rotated input design approach, the matrix U_f may show nearly collinear rows, so that the overall matrix Z_f may be again near to rank deficiency, because some of mr singular values associated to U_f may not be significant. That is, although the rotated input design approach may be useful to excite the system in weaker directions, the strong alignment of the inputs in specific directions may constitute a problem in all subspace identification algorithms which project input data as well as output data.

4. Case studies

4.1. Introduction

Two different case studies are presented to compare the behavior of rotated inputs vs. that of random OL inputs and CL inputs obtained with random setpoint signals. OL identification data are constructed by using as inputs generalized binary noise (GBN) signals, either uncorrelated or correlated in the case of rotated inputs, as outlined at end of Section 3.1. Introduced by Zhu [13], GBN signals have many favorable features, in particular in terms of frequency content, which is typically superior to that of pseudo-random binary noise (PRBS) and of step signals. Closed-loop data are obtained by using an MPC regulator as that described in [31], based on a preliminary (erroneous) model of the system to be identified. In CL data collection, setpoints for the controlled variables are GBN signals. These simulations are performed in order to find the most appropriate test input design for subspace identification of ill-conditioned processes. The identified models are compared in terms of frequency responses and in terms of performance of (unconstrained) MPC regulators based on the identified models. The final goal of these studies is to explain why CL data collection is to be preferred for SID of ill-conditioned plants.

Example 1. This is the "classical" two input - two output high-purity distillation column studied by Skogestad and Morari [32]. A transfer function model of the process is the following:

[2] Notice that the adopted identification algorithm performs an SVD of the projected data matrix Z, defined in (8). However, it is obvious from (8) that rank $Z \leqslant$ rank Z_f, so that the rank bound on Z_f is a valid bound for Z as well.

Table 1
Inputs and outputs for Example 2

Input		Output	
u_1	Pressure	y_1	Distillate composition
u_2	Reboiler heat duty	y_2	Bottom composition
u_3	Reflux rate	y_3	Valve opening
–	–	y_4	Valve opening
–	–	y_5	Valve opening

$$y = \begin{bmatrix} \frac{87.8}{194s+1} & \frac{-87.8}{194s+1} + \frac{1.4}{15s+1} \\ \frac{108.2}{194s+1} & \frac{-108.2}{194s+1} + \frac{-1.4}{15s+1} \end{bmatrix} u, \tag{21}$$

in which the outputs are logarithmic distillate purity and logarithmic bottom impurity, the inputs are reflux and boil-up rates and the time constants are in minutes. Normally, distributed output noise with a noise-to-signal ratio of 0.10 is added to both outputs, and a sampling time of 5 min is considered. CL identification data are obtained by using an MPC regulator, based on the following "erroneous" model:

$$y = \begin{bmatrix} \frac{75}{180s+1} & \frac{-75}{180s+1} + \frac{1.6}{19s+1} \\ \frac{105}{180s+1} & \frac{-105}{180s+1} + \frac{-1.6}{19s+1} \end{bmatrix} u. \tag{22}$$

Example 2. This is a 3×5 linear system that represents a high-purity distillation column for the separation benzene–toluene (see Table 1 for a detailed list of inputs and outputs). The condition number of its gain matrix is 140. This linear discrete-time system, omitted in the sake of space, is a representation of the rigorous non-linear simulation model, developed in Octave.[3] A 5 min time delay (i.e. 5 samples) is considered on head and bottom impurity outputs, and normally distributed output noise with a noise-to-signal ratio of 0.10 is added to all outputs. CL identification data are collected by using an MPC regulator based on a preliminary model, identified via step tests on each input variable independently. For both examples, all datasets contain 2500 samples.

Model validation is conducted, in both examples, by means of a frequency error analysis, as described. Defining $G^r(z)$ and $G^{id}(z)$ as the real and identified discrete transfer function models, respectively, the following scalar parameter is defined to measure the quality of the identified model:

$$\bar{\epsilon} = \frac{\|G^{id}(z) - G^r(z)\|_\infty}{\|G^r(z)\|_\infty} \triangleq \sup_{\omega>0} \underbrace{\left[\frac{\sigma_1\left(G^{id}(e^{i\omega T_s}) - G^r(e^{i\omega T_s})\right)}{\sup_{\omega>0}\sigma_1(G^r(e^{i\omega T_s}))}\right]}_{\epsilon(\omega)} \tag{23}$$

in which $\|\cdot\|_\infty$ denotes the $\mathscr{H}_\infty$ norm (in discrete-time sense), T_s is the sampling time, and $\sigma_1(\cdot)$ denotes the largest singular value of $(\cdot)$. It is clear from (23) that the lower $\bar{\epsilon}$ the better the identified model.

[3] Octave is freely available at http://www.octave.org.

4.2. Inputs and outputs plots (Example 1)

Fig. 1 shows the input sequences for three different data collection schemes: OL data with random inputs, OL data with rotated inputs and CL data with random setpoints. For the moment, the exact system gain matrix is used to generate the sequence of rotated inputs, according to the procedure outlined in Section 3.1. The effect of an incorrect rotation angle is discussed later in Section 4.6. Fig. 2 reports the inputs in the plane (u_1, u_2) and the outputs in the plane (y_1, y_2) obtained from the three data collection

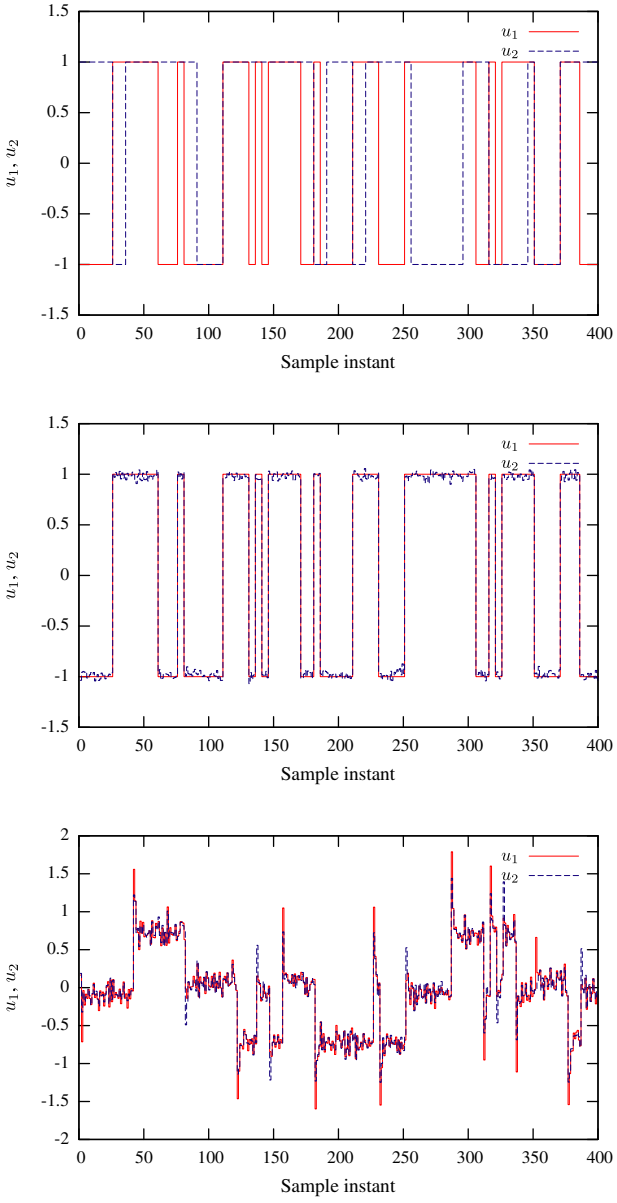

Fig. 1. Example 1: standard GBN input (top), rotated GBN input (middle) and closed-loop input (bottom) signals.

schemes. It is possible to see that in case of random inputs, the outputs are placed in a narrow region with a dominant direction, which is the high-gain direction. Thus, it is clear that information from the low-gain direction is "hidden", because it is much smaller in magnitude and it cannot be evaluated correctly from the dataset, because of the presence of noise. In case of rotated inputs, output data are contained in a region where all the directions are of similar magnitude, and this shows that the rotated input design reached its goal. However, it is possible that inputs are not enough informative, because they are strongly aligned in a specific direction. Finally, the bottom plots show that outputs obtained in closed-loop present the same advantages of those coming from rotated inputs, i.e. they are not aligned in a preferred direction. At the same time, it is possible to see that the inputs obtained in closed-loop are not strongly aligned in a preferred direction either.

4.3. Model order recovering (Example 1)

In this section, order recovering results are shown and discussed for Example 1. The past and future horizon is

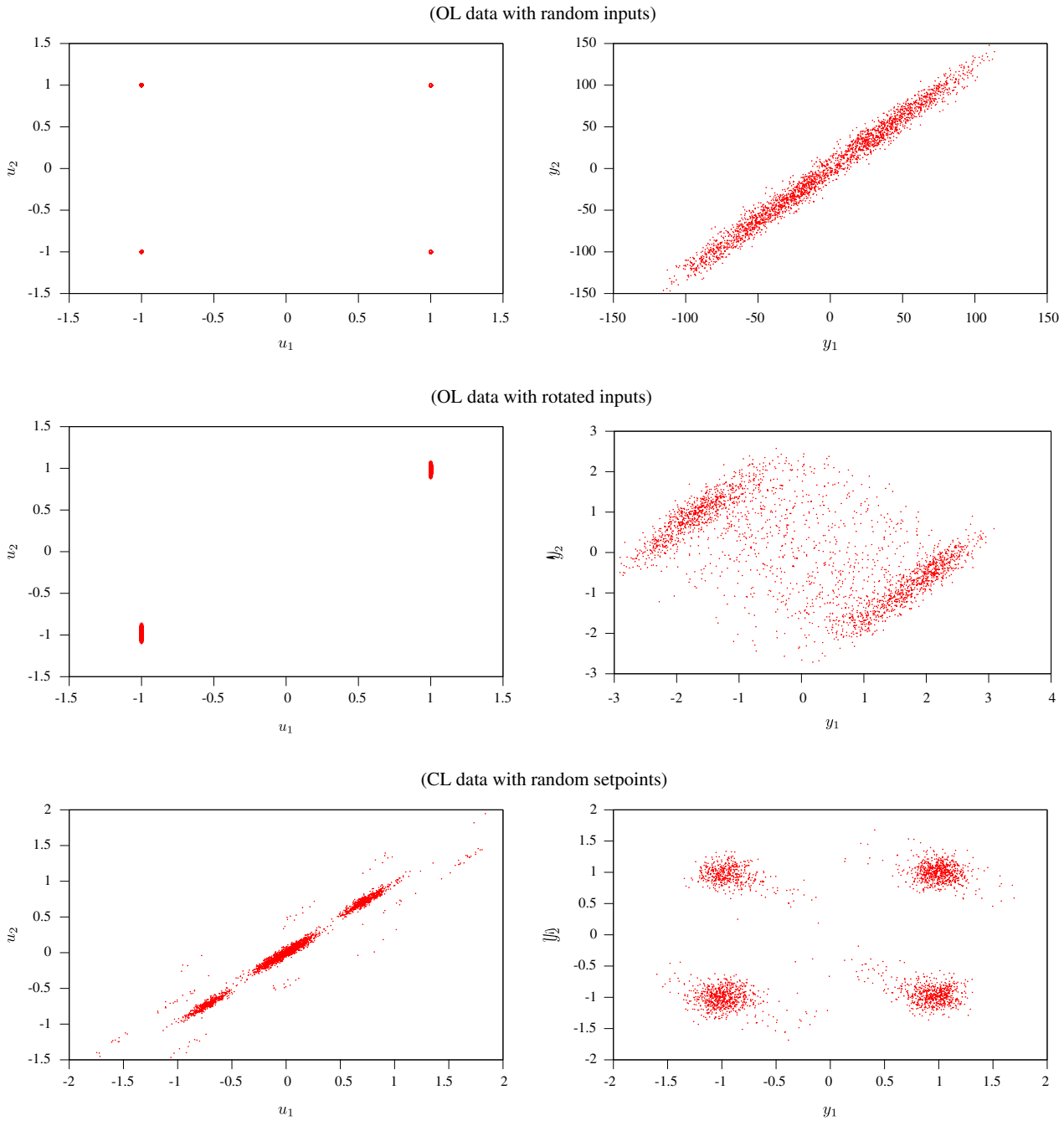

Fig. 2. Example 1: alignment of inputs (left) and outputs (right) in three cases: OL data collection with random inputs, OL data collection with rotated inputs, CL data collection with random setpoints.

equal to $r = 30$, while a value of 0.05 is assumed for ρ in (10) to select the system order from the singular values of Z. During this work it was experienced that open-loop rotated inputs generate mistakes in order determination when used with subspace algorithms which project not only the future outputs matrix but also the future inputs matrix (such as the one adopted in this paper). This occurs because the matrix U_f shows nearly collinear rows given the strong alignment of the inputs in a fixed direction as shown in Fig. 2.

Singular values of projected matrix Z are computed in three different cases: OL data collection using uncorrelated GBN inputs, OL data collection using rotated GBN inputs, and CL data collection using uncorrelated GBN setpoints. For Example 1, it results that neither OL data obtained from uncorrelated inputs nor that obtained from rotated inputs allow the projection algorithm to compute the correct model order (which is 2): they give a model order of 5 and a model order of 12, respectively. On the other hand CL data permit the algorithm to recover the correct model order. In order to confirm that OL rotated inputs may work well with particular SID algorithms, it was verified that using Misra and Nikolaou's algorithm [19] on the same rotated inputs dataset leads to the correct model order. Notice that if OL uncorrelated inputs data are used, Misra and Nikolaou's identification method computes a model order of 1.

For Example 2, the orders of the models identified with the orthogonal projection method are 17 in case of OL random inputs, 17 in case of rotated OL inputs and 14 in case of CL random inputs, respectively. Notice that the correct order is 24 (including ten state variables for outputs delay).

4.4. Quality of the identified models

In this section, quality of the identified models is evaluated and compared in terms of frequency responses. Fig. 3 shows the relative error $\epsilon(\omega)$ vs. frequency of three models identified from different datasets (OL with random inputs, OL with rotated inputs and CL with random setpoints). The top plot concerns with Example 1 and the bottom plot concerns with Example 2. From these results the superior quality of the model identified from closed-loop data clearly appears.

It is well known that identification with subspace methods may suffer from sensitivity to different noise realizations. For this reason, a Monte-Carlo study of 600 simulations was performed on Example 1: 200 simulations are conducted in OL using uncorrelated GBN input signals, 200 simulations are conducted with OL rotated GBN input signals, and 200 simulations are conducted in CL with uncorrelated GBN setpoints. The results of this study are reported in Fig. 4, in which the density function of the parameter $\bar{\epsilon}$ for the models identified from the three datasets is depicted. It can be seen that the curves do not follow a normal (symmetric) distribution and are right-skewed. Mean values are also reported in the plot, and it is possible to see that models obtained from CL data have superior accuracy (i.e. lower mean of $\bar{\epsilon}$). Moreover, it is possible to see that models obtained from CL data are more precise (i.e. lower variance of $\bar{\epsilon}$). In conclusion, this Monte-Carlo simulation study shows that CL data give superior models in terms of accuracy and precision compared to data obtained by OL random inputs and by OL rotated inputs.

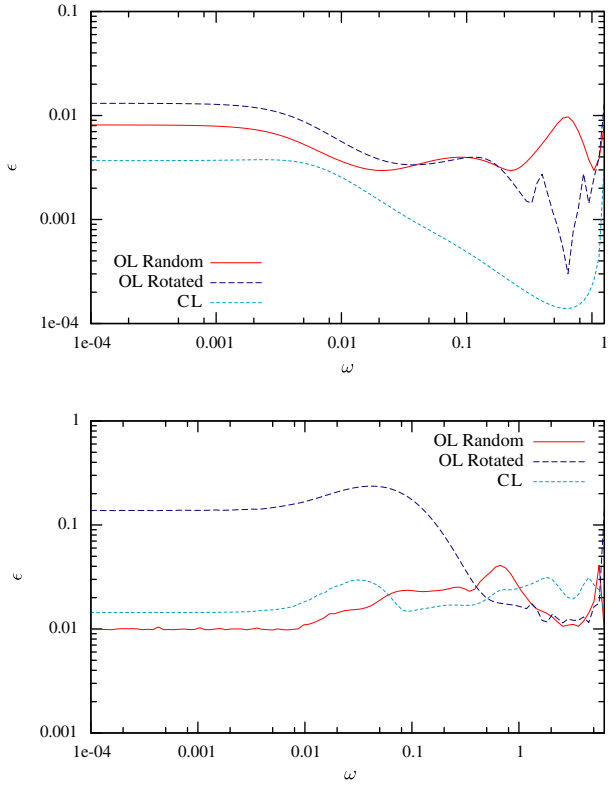

Fig. 3. Model error parameter $\epsilon(\omega)$ vs. frequency for three identified models: Example 1 (top) and Example 2 (bottom).

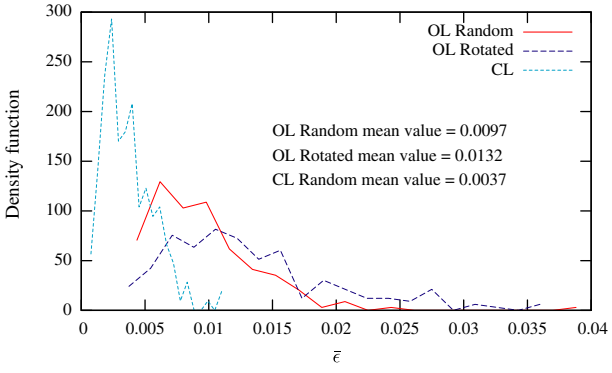

Fig. 4. Example 1: density function of $\bar{\epsilon}$ obtained from a Monte-Carlo study.

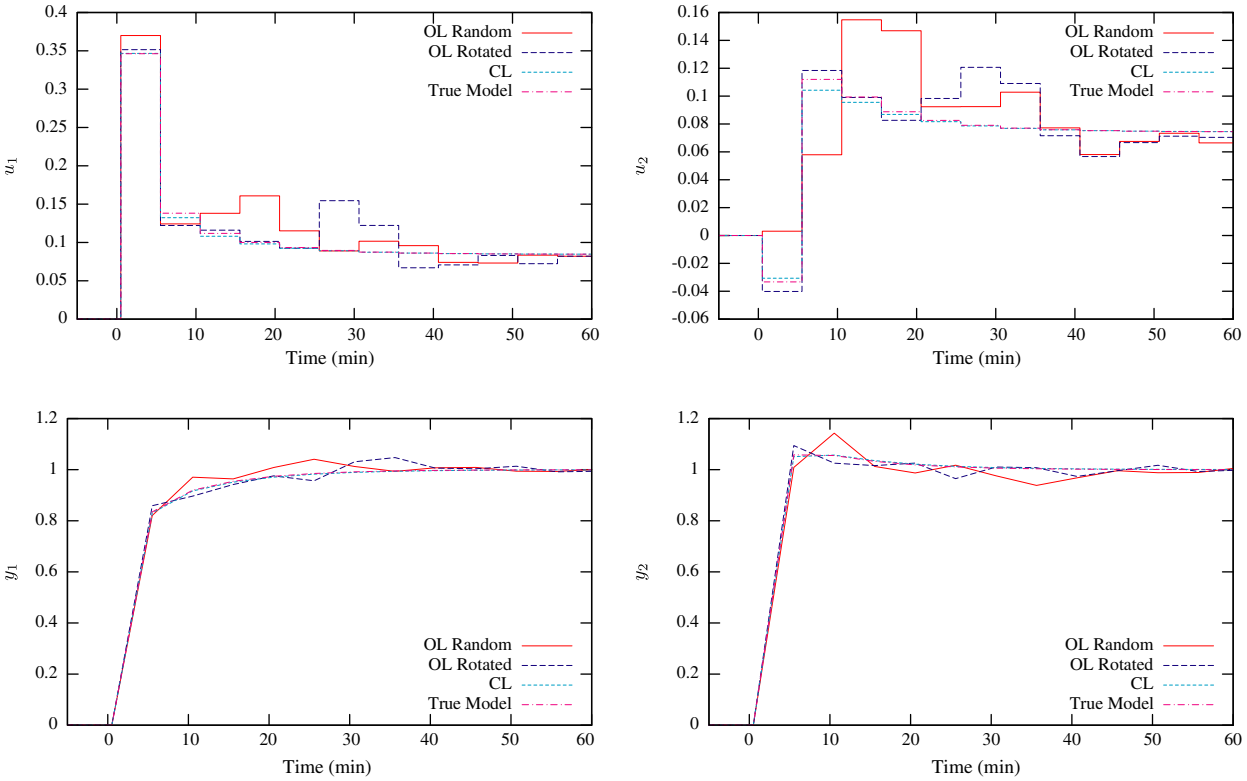

Fig. 5. Example 1: closed-loop inputs and outputs during a setpoint change in the direction $[1,1]^T$.

4.5. Effect of the identified models on MPC closed-loop behavior

The closed-loop behavior of several MPC regulators, equivalent in all tuning parameters (omitted in the sake of space) but based on different models, is compared. Three MPCs are designed on the models identified from OL random, OL rotated and CL inputs, respectively; a fourth MPC is designed on the true model (21).

For Example 1, closed-loop inputs and outputs obtained for a setpoint change imposed at time 0, are shown in Fig. 5. It is clear that the regulator based on the model obtained from CL data shows superior performance (essentially equal to that of the regulator based on the true model). The regulators based on models identified from OL random and OL rotated inputs, instead, show a worse performance especially in terms of relevant fluctuations of the manipulated variables. These results are in definite agreement with the model error frequency responses shown in Fig. 3.

Fig. 6 shows closed-loop inputs and the first three outputs of Example 2, during a setpoint change on the head and bottom impurity outputs. Also for this case the regulator based on the model obtained from CL signals performs well, slightly better than the regulator based on the model identified from OL random signals (which shows a larger settling time in the output variables). For this case, the MPC based on the model identified from OL rotated signals shows poor performance with large oscillations of outputs and inputs. This poor behavior is a clear consequence of the large model error, especially at low frequency, shown in Fig. 3.

From these simulation results, as well as from other case studies, it is possible to state that CL data collection is to be preferred for ill-conditioned processes, both in the sense of model error frequency response, and in terms of superior closed-loop performance when the identified model is used inside an MPC regulator. OL data collection, even with a rotated input approach, is instead to be avoided for subspace identification of ill-conditioned processes.

These evidences can be explained, with the aid of the plots shown in Fig. 2, as follows. CL data collection with random setpoints forces the outputs and the inputs of the system in many different directions, i.e. both in high-gain and low-gain directions. On the other hand, OL data collection with random inputs forces the outputs to be aligned mostly in high-gain directions, whereas OL data collection with rotated inputs certainly equalizes the contribution of high and low gain directions, but the corresponding inputs may be too strongly aligned in specific directions. This shortcoming is particularly relevant when subspace algorithms projecting the input data matrix (along with the output data matrix) are applied. Other disadvantages of rotated inputs are discussed in the next paragraph.

4.6. Practical issues in implementing rotated inputs

Rotated inputs may guarantee good performances in model order recovery when used with specific identification methods (such as that in [19]). However, until now it has

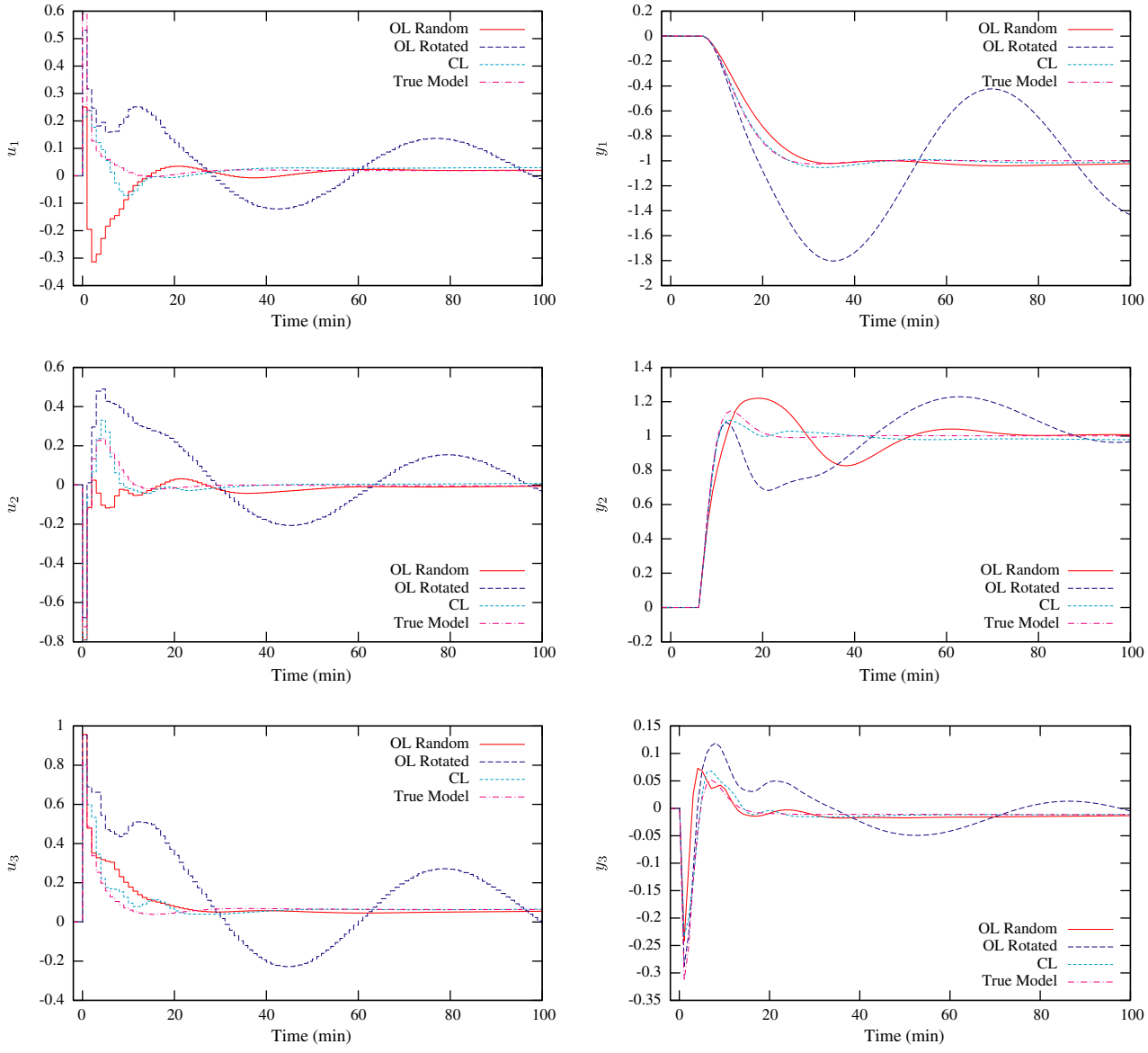

Fig. 6. Example 2: closed-loop inputs and three outputs during a setpoint change on the product compositions.

not been shown the sensitivity of this method to errors in the rotation angle, that is how rotated inputs work when the selected rotation angle is different from the optimal rotation angle, computed via SVD of the "exact" gain matrix. When the rotation angle is $\pi/3$ rather than about $\pi/4$ that is obtained from SVD of the exact system gain matrix, the model order identified from Misra and Nikolaou's algorithm [19] is 1. This confirms that rotated inputs may work well only if the correct rotation angle is applied. In practice, since the true model order is not known, it is quite difficult to tell from the singular values whether the applied rotation angle is correct or not.

5. Conclusions

Ill-conditioned processes are difficult to be identified from data, because of problems that derive mainly from the lack of information in the weaker directions of the system. A possible solution, proposed in the literature and deeply discussed throughout this paper, is the use of tailored inputs known as "rotated" inputs [11,19], which excite the system equally in high-gain and low-gain directions. To construct these inputs, the system gain matrix (which is however not known precisely) can be decomposed via singular value decomposition, and inputs are designed to generate output signals of the same magnitude both in weak and in strong directions. In this work this input design approach was generalized to non-square multivariable systems of arbitrary dimension and applied to two case studies of ill-conditioned processes. Results clearly show that, dealing with an open-loop data collection scheme, rotated inputs may grant better models than uncorrelated inputs, but these superior results are strongly dependent on the subspace identification algorithm used. In particular rotated inputs appear inappropriate for the use with the orthogonal projection method [7]. Moreover,

the effectiveness of rotated inputs is strongly related to the accuracy of the applied rotation angle(s), which in general must be found by trial and error. Results show that closed-loop data collection, instead, guarantees superior models, both in terms of lower error in frequency response between the identified models and true process, and most importantly in terms of higher performance achieved by model predictive controllers based on the identified models. Furthermore, this closed-loop test design is to be preferred, because random setpoints can be easily generated without necessity of several trials to find the most appropriate rotation angle(s) of the inputs.

References

[1] L. Ljung, System Identification: Theory For The User, Prentice & Hall, 1999.

[2] P. Van Overschee, B. De Moor, N4SID: Subspace algorithms for the identification of combined deterministic stochastic systems, Automatica 30 (1994) 75–93.

[3] M. Verhaegen, Identification of the deterministic part of MIMO state space models given in innovation form from input–output data, Automatica 30 (1994) 61–74.

[4] P. Van Overschee, B. De Moor, A unifying theorem for three subspace system identification algorithms, Automatica 31 (1995) 1853–1864.

[5] W.E. Larimore, Automated and optimal system identification by canonical variables, in: American Control Conference, Baltimore, MD, vol. 2, IEEE, Piscataway, NJ, 1994, pp. 1640–1644.

[6] J. Wang, S.J. Qin, A new subspace identification approach based on principal component analysis, J. Proc. Cont. 12 (2002) 841–855.

[7] B. Huang, S.X. Ding, S.J. Qin, Closed-loop subspace identification: an orthogonal projection approach, J. Proc. Cont. 15 (2005) 53–66.

[8] Y. Zhu, Multivariable process identification for MPC: the asymptotic method and its applications, J. Proc. Cont. 8 (2) (1998) 101–115.

[9] M. Gevers, L. Ljung, Optimal experiment designs with respect to the intended model application, Automatica 22 (1986) 543–554.

[10] C.W. Koung, J.F. MacGregor, Design of identification experiments for robust control. A geometric approach for bivariate processes, Ind. Eng. Chem. Res. 32 (1993) 1658–1666.

[11] C.W. Koung, J.F. MacGregor, Identification for robust multivariable control: the design of experiments, Automatica 30 (1994) 1541–1554.

[12] H. Hjalmarsson, M. Gevers, F. de Bruyne, For model-based control design, closed-loop identification gives better performance, Automatica 32 (1996) 1659–1673.

[13] Y. Zhu, Multivariable System Identification for Process Control, Elsevier Science, 2001.

[14] S. Skogestad, M. Morari, Implications of large RGA-elements on control performance, Ind. Eng. Chem. Res. 26 (1987) 2323–2330.

[15] G. Pannocchia, Robust disturbance modeling for model predictive control with application to multivariable ill-conditioned processes, J. Proc. Cont. 13 (2003) 693–701.

[16] E.W. Jacobsen, S. Skogestad, Inconsistencies in dynamic models for ill-conditioned plants: Application to low-order models of distillation columns, Ind. Eng. Chem. Res. 33 (1994) 631–640.

[17] K.E. Häggblom, J.M. Böling, Multimodel identification for control of an ill-conditioned distillation column, J. Proc. Cont. 8 (2) (1998) 209–218.

[18] A. Chiuso, G. Picci, On the ill-conditioning of subspace identification with inputs, Automatica 40 (2004) 575–589.

[19] P. Misra, M. Nikolaou, Input design for model order determination in subspace identification, AIChE J. 49 (8) (2003) 2124–2132.

[20] J.S. Conner, D.E. Seborg, An evaluation of MIMO input designs for process identification, Ind. Eng. Chem. Res. 43 (2004) 3847–3854.

[21] P. Stec, Y. Zhu, Some study on identification of ill-conditioned processes for control, in: Proceedings of the American Control Conference, Arlington, VA, USA, 2001, pp. 1202–1207.

[22] B.L. Cooley, J.H. Lee, Control-relevant experiment design for multivariable systems described by expansions in orthonormal bases, Automatica 37 (2001) 273–281.

[23] R.B. Gopaluni, R.S. Patwardhan, S.L. Shah, Experiment design for MPC relevant identification, in: Proceedings of American Control Conference, Anchorage, AK, USA, 2002, pp. 2713–2718.

[24] Y. Zhu, P. Stec, Simple control-relevant identification test methods for a class of ill-conditioned processes, J. Proc. Cont. 16 (2006) 1113–1120.

[25] H. Hjalmarsson, From experiment design to closed-loop control, Automatica 41 (2005) 393–438.

[26] M.J. Bruwer, J.F. MacGregor, Robust multi-variable identification: optimal experimental design with constraints, J. Proc. Cont. 16 (2006) 581–600.

[27] Q. Zhan, T. Li, C. Georgakis, Steady state optimal test signal design for multivariable model based control, Ind. Eng. Chem. Res. 45 (2006) 8514–8527.

[28] S.J. Qin, W. Lin, L. Ljung, A novel subspace identification approach with enforced causal models, Automatica 41 (2005) 2043–2053.

[29] G. Pannocchia, A. Micchi, R. Bulleri, A. Brambilla, G. Marchetti, Multivariable subspace identification and predictive control of a heat-integrated superfractionator rigorous model, in: Proceedings of the ADCHEM 2006 International Symposium on Advanced Control of Chemical Process, vol. I, Gramado, Brazil, 2006, pp. 421–426.

[30] L. Ljung, T. McKelvey, Subspace identification from closed-loop data, Signal Process. 52 (1996) 209–215.

[31] G. Pannocchia, J.B. Rawlings, Disturbance models for offset-free model predictive control, AIChE J. 49 (2003) 426–437.

[32] S. Skogestad, M. Morari, LV-control of a high purity distillation column, Chem. Eng. Sci. 43 (1988) 33–48.