Detecting Stationary Gain Changes in Large Process Systems

Germán A. Bustos [a] , Alejandro H. González [a] & Jacinto L. Marchetti [a]

[a] INTEC (CONICET and Universidad Nacional del Litoral) , Santa Fe , Argentina
Accepted author version posted online: 09 Dec 2013.Published online: 09 Dec 2013.

PLEASE SCROLL DOWN FOR ARTICLE

ACCEPTED MANUSCRIPT

Detecting Stationary Gain Changes in Large Process Systems

Germán A. Bustos[1], Alejandro H. González[1], Jacinto L. Marchetti[1,]

[1]INTEC (CONICET and Universidad Nacional del Litoral), Santa Fe, Argentina

Address correspodence to Jacinto L. Marchetti, INTEC (CONICET and Universidad Nacional del Litoral), Güemes 3450 (3000), Santa Fe, Argentina E-mail: jlmarch@santafe-conicet.gov.ar

Abstract

Stationary process gains are critical model parameters for determining targets in commercial MPC technologies. Consequently, important savings can be reached by accessing an early prevention method capable of detecting whether the actual process moves away from the modeled dynamics, particularly by indicating when the process gains are no longer represented by those included in the model identified during commissioning stages. In this first approach, a subspace identification method is used under open-loop process condition to estimate the process gain matrix. The main reason for using the subspace identification (SID) method is that it works directly with raw data; it directly yields a multivariable state space model and has proved to be successful in dealing with multivariable processes and periodic batch-wise data collection. To detect significant changes in the estimator population, a monitoring sequence of hypothesis tests can be done through simple confidence limits directly on each gain estimator, or increasing the sensitivity by using the exponentially weighted moving average (EWMA) or the cumulative sum (CUSUM) algorithms. The objective of this aticle is to present a rational combination of inferential tools capable of detecting which gain of a multivariable model starts moving away from its original value. The anticipated knowledge of these events could provide a warning to process engineers and prevent targeting process conditions with wrong gain estimations. The regular follow-up of the gain matrix should also help to localize those dynamics needing an updating identification and reduce the frequency of time-consuming re-identification of the complete model.

KEYWORDS: LP-MPC; Multivariable processes; Steady-state gains; Subspace

identification

INTRODUCTION

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Model predictive control (MPC) has wide application in the chemical process industry and other industrial sectors (González et al., 2006; Qin and Badgwell, 2003). Commercial MPC systems are typically implemented in conjunction with a steady-state linear (LP) or quadratic programming (QP) optimizers (Ying and Joseph, 1999), whose main function is to track the economic optimum and provide feasible set-points or targets to the predictive controller. However, despite the widespread adoption of these two-level control systems, occurrences of poor performance have been reported. There are frequent claims that model mismatches push the operation away from the real optimum, and large variations in the computed targets have been observed (Nikandrov and Swartz, 2009). Since the stationary process gains are critical model parameters to determine MPC targets, important savings can be obtained by accessing an early prevention system capable of indicating when the actual process moves away from the modeled behavior, particularly by indicating when the actual process gains are no longer represented by those included in the model identified during early commissioning stages.

Modern data acquisition has allowed process control systems to become multivariable and provided the technological base to develop monitoring applications capable of a simultaneous surveillance of several correlated characteristic variables. This fact has motivated the challenge of extending several single-variable statistical methods to multivariate applications. However, in spite of the important advances made in monitoring multivariable systems, we adopt in this case a multiple single-gain strategy. The reason for this choice is in line with the main purpose of this work, which is to detect and localize those process gains that have moved away from the expected model values

ACCEPTED MANUSCRIPT

and, whenever possible, to estimate the amplitude of the individual changes. Notice that this goal is quite different from the detection of unexpected changes in the operating conditions or the prediction of multivariable measurement systems.

This work presents the first results of a methodology suitable for the estimation of multivariable process gains directly from raw data and creates the basis for a detection tool capable of providing warning signals when significant gain changes have occurred. In pursuing this objective, we have observed increased attention to subspace identification (SID) methods for industrial applications over the past decade. An advantage of the subspace method is that it directly yields a multivariable state space model and considers the correlation of the output measurements in the identification, thus leading to more accurate models. Compared to finite step response (FIR) models, experience has shown that SID leads to more accurate estimates of gain and gain ratios, which are critical to capturing the true degrees of freedom in MPC and ensuring reliable LP performance (Darby and Nikolaou, 2012).

A typical difficulty arising when applying empirical modeling technologies to real processes is the need of artificially disturbing the process and finding operating points under which the signal-to-noise ratio is good enough to perform the desired identification, something that can be a distressing exercise for process operators too. Since this problem remains unsolved, the strategy analyzed here for detecting changes assumes the viability of using periods of time to persistently excite the system. Hence, a

sequence of batch-wise data collections is considered as relevant information for the follow-up procedure shown here.

Another source of concern, particularly in the chemical process industry, is the presence of capacities or accumulators that generate dynamic behaviors very closely described by integral modes, which preclude the open-loop application of standard pseudo-random binary signals (PRBS). To overcome this difficulty, a rational approach based on a modified PRBS and a specific algebraic treatment is proposed in this work.

The article is organized as follows: after a short introduction on the motivations and goals of this project, a brief discussion of the SID method is given, which the reader can complement with the material in Appendices A and B. Then, based on the estimated model matrices, the gain-matrix estimator is defined, where the emphasis is on the different treatment required when there are dynamic evidences of integral-mode behavior. The next section summarizes the fundamentals of well-established techniques coming mostly from statistical quality control (SQC), like the exponentially weighted moving average (EWMA) and the cumulative sum (CUSUM), that are successfully used here for assessing the occurrence of individual gain changes. Then, after presenting a 5×3 linear system as a demonstration case, a preliminary analysis is made for determining the convenient dimensions of the block Hankel matrices required by the SID method, as well as the overall model order to be used. Some simulation results are presented, and, finally, conclusions are given, together with comments about future work associated with this subject.

A part from the specific treatment given to processes with integral mode dynamics, the novelty in this article cannot be found in the individual techniques SID, CUSUM, or EWMA, but in the integration of such techniques. To the authors' knowledge, no attempt has been made before to successfully combine these concepts to provide an effective solution for monitoring the gains of multivariable systems, perhaps because they come from different engineering areas.

SUBSPACE IDENTIFICATION (SID) METHOD

Early in the 1990s, a new identification method for dynamic systems received the attention of many academics and practitioners. The subspace identification (SID) method has the appealing feature of allowing the direct use of raw data with low preprocessing needs and with applicability to multivariable process systems. Several analyses and applications have been reported since then; some of the more cited are Van Overschee and De Moor (1996), Favoreel et al. (2000), and Katayama (2005).

Most subspace approaches fall into the so-called unification theorem proposed by Van Overschee and De Moor (1996), the three best known are N4SID (Van Overschee and De Moor, 1994), canonical variate analysis (CVA; Larimore, 1990), and multivariable output error state space (MOESP; Verhaegen and Dewilde, 1992). They provide reliable state-space models of multivariable linear time invariant (LTI) systems directly from input-output data and do not require iterative optimization procedures; this basically means that there are no problems of local minima, convergence, or initialization.

According to the goal of this work, the main task consists in the online estimation of the steady-state gains of multiple-input/multiple-output (MIMO) systems using a standard subspace approach (Van Overschee and De Moor, 1996). When using this method, the specific model matrices are obtained from projections of the subspaces generated by the input and output data, which are collected under persistent excitation. The subspace projections implicit in the method lead to capturing most of the dynamic information from the data, and at the same time they contribute to removing noisy components. The notation used in this article follows that commonly used in the extensive literature available on SID methods (Van Overschee and De Moor, 1996). Please note the convenience of reading Appendix A before proceding to the material dealing with the details of the SID method in Appendix B.

The SID method applied here considers the following state-space model:

$$x(k+1) = A\,x(k) + B\,u(k) \tag{1}$$

$$y(k) = C\,x(k) + D\,u(k) + e(k) \tag{2}$$

where $x(k) \in \mathbb{R}^n$ stands for a n-dimensional state, $u(k) \in \mathbb{R}^m$ represents the m inputs to the system, $y(k) \in \mathbb{R}^l$ is the l-dimensional output, and $e(k) \in \mathbb{R}^l$ <AQ: please check symbols that did not translate>is a noise depending on nonassignable common causes of variation and measurement errors. Notice the convenience of taking the sampling interval $T = (k+1)T - kT$ as equal to the sampling interval used by the LP-MPC system to be monitored.

The problem solved by the SID method can be summarized as follows: given a large enough data set $\{y_k, u_k\}$ obtained from an unknown system, find the estimated matrices $\hat{A}$, $\hat{B}$, $\hat{C}$, and $\hat{D}$ of the state-space representation given by Equations (1) and (2).

GAIN MATRIX ESTIMATION

Stable Processes With Or Without Dead Times

Once the estimations of the model matrices **A**, **B**, **C**, and **D** are available, the stationary condition predicted by the model has to satisfy the following relationships:

$$x_{SS} = A x_{SS} + B u_{SS} \tag{3}$$

$$y_{SS} = C x_{SS} + D u_{SS} \tag{4}$$

where x_{SS}, u_{SS}, and y_{SS} stand for stationary values of the state and the input and the output variables, respectively. Substituting x_{SS} from Equation (3) into Equation (4) and rearranging the equations, the system gain matrix is determined by the relationship between the stationary input u_{SS} and the stationary output y_{SS}:

$$y_{SS} = \left[C(I - A)^{-1} B + D \right] u_{SS} = G u_{SS} \tag{5}$$

Hence, assuming that an identification method provides the estimations $\hat{A}$, $\hat{B}$, $\hat{C}$, and $\hat{D}$, the gain matrix can be estimated by

$$\hat{G} \triangleq \hat{C}\left(I - \hat{A}\right)^{-1} \hat{B} + \hat{D} \tag{6}$$

Notice that despite the fact that the gains can be estimated from different steady states, i.e., without transient information, in many practical cases is not possible to wait until the stationary conditions are reached. Furthermore, there are many real processes where the

stationary states are just occasional. The SID method's capacity for working with transient data is essential when the final goal—to be developed in future work—is gain estimation during closed-loop operation with constraints. Besides, real process gains are not expected to change from one sampling time to the next one, i.e., during a 1 or 2 min interval; these are typically slow changes that permit the monitoring task to be executed no more frequently than once a day, thus giving enough time for periodic batch-wise data collection.

Procedure When There Are Integrating Modes In The Process

For input-output relationships containing integrating modes (i.e., poles at the origin or, equivalently, unitary eigenvalues of matrix **A**) the concept of system gain must be revised. In fact, in a pure integrating system, the instantaneous value of the output does not depend on the instantaneous value of the input, but on the "history" of the input behavior; more precisely, the output depends on the integral of the complete behavior of the input. This clearly implies that, at the stationary state, there is no longer a linear fixed relationship between the stationary output and the stationary input, and, thus, the concept of system gain, in the sense defined in the preceding section, is no longer applicable. Algebraically, the presence of an integrating system can be detected by the impossibility of inverting matrix $(I - A)$ to obtain **G**, as in Equation (6).

Let us assume a general MIMO system, with $D = 0$ where every transfer function is stable except $G_{ji}(s)$, which includes an integrating mode between u_i and y_j:

$$
\begin{bmatrix} y_1(s) \\ \vdots \\ y_j(s) \\ \vdots \\ y_l(s) \end{bmatrix} = \begin{bmatrix} G_{11}(s) & \cdots & & G_{1i}(s) & & \cdots & G_{1m}(s) \\ \vdots & \ddots & & \vdots & & \ddots & \vdots \\ G_{j1}(s) & \cdots & G_{ji}(s) = \dfrac{K_{ji}}{s}\tilde{G}_{ji}(s) & & \cdots & G_{jm}(s) \\ \vdots & \ddots & & \vdots & & \ddots & \vdots \\ G_{1l}(s) & \cdots & & G_{li}(s) & & \cdots & G_{lm}(s) \end{bmatrix} \begin{bmatrix} u_1(s) \\ \vdots \\ u_i(s) \\ \vdots \\ u_m(s) \end{bmatrix} \qquad (7)
$$

Following the above comments, note that K_{ji} is not a stationary gain as those found in he all-stable-modes case, but the final slope of the step response of the transfer function $G_{ji}(s)$. To analyze this case, let us start by considering the diagonal version of the state transition matrix $\mathbf{A}$, A_d, of the discrete state-space model, which can be obtained by a Jordan decomposition:

$$
A_d = V^{-1}AV = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \in \mathbb{R}^{n \times n} \qquad (8)
$$

where $\lambda_1, \lambda_2, ..., \lambda_n$, are the stable modes of the system, $\lambda_1 = 1$ represents the integrating mode, and V is the invertible transformation matrix. Furthermore, the corresponding transformed matrices $\mathbf{B_d}$ and $\mathbf{C_d}$ can be obtained as:

$$
B_d = V^{-1}B = \begin{bmatrix} 0 & \cdots & B_{d,1i} & \cdots & 0 \\ B_{d,21} & \cdots & B_{d,2i} & \cdots & B_{d,2m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{d,n1} & \cdots & B_{d,ni} & \cdots & B_{d,nm} \end{bmatrix} \in \mathbb{R}^{n \times m}, \; C_d = CV = \begin{bmatrix} 0 & C_{d,12} & \cdots & C_{d,1n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{d,j1} & C_{d,j2} & \cdots & C_{d,jn} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & C_{d,l2} & \cdots & C_{d,ln} \end{bmatrix} \in \mathbb{R}^{l \times n}
$$

(9)

Once the representation is in this form, the integrating modes can be separated from the

model by removing the parameters exclusively associated with the integrating modes,

obtaining in this way a stable system. That is, let us consider the following matrices:

$$\tilde{A}_d = \begin{bmatrix} \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}, \ \tilde{B}_d = \begin{bmatrix} B_{d,21} & \cdots & B_{d,2m} \\ \vdots & \ddots & \vdots \\ B_{d,n1} & \cdots & B_{d,nm} \end{bmatrix} \in \mathbb{R}^{(n-1)\times m},$$

$$\tilde{C}_d = \begin{bmatrix} C_{d,12} & \cdots & C_{d,1n} \\ \vdots & \ddots & \vdots \\ C_{d,l2} & \cdots & C_{d,ln} \end{bmatrix} \in \mathbb{R}^{l\times(n-1)}$$

where matrix $\tilde{A}_d$ is obtained by elimination of the line and the column associated with λ_1

$= 1$, while $\tilde{B}_d$ and $\tilde{C}_d$ are obtained by removing the line and the column associated,

respectively, with the integrating mode. The new system ($\tilde{A}_d$, $\tilde{B}_d$, $\tilde{C}_d$) is completely

stable, and the gain matrix of this reduced system is given by

$$G = \tilde{C}_d \left(I - \tilde{A}_d \right)^{-1} \tilde{B}_d \in \mathbb{R}^{l\times m} \tag{10}$$

Notice that this matrix **G** has the same dimension that the original system. The "gain"

corresponding to the integrating mode is zero, which should be interpreted as a transfer

function without stationary gain.

The diagonal form of (A_d, B_d, C_d) also provides a simple way to identify the transfer

functions that include an integrating mode and to find the values of the corresponding

integration constant, K_{ji}. As shown by Equation (9), matrix B_d has a single non-null

element $B_{d,1i}$ in the row corresponding to the integrating mode, while matrix C_d presents a single non-null element $C_{d,j1}$ in the column corresponding to the integrating mode. This means that the input associated to the integrating mode is "i," and the associated output is "j," i.e., the transfer function with the integrating mode is $G_{ji}(s)$, and there is no need to know this fact a priori. Furthermore, integrating the first state difference equation,

$$x_1(k+1) = x_1(k) + B_{d,1i}u_i \tag{11}$$

and using the output in Equation (2) with $D = 0$ permits the constant associated to the integrating mode to be estimated by

$$\hat{K}_{ji} = \frac{1}{T}\hat{C}_{d,j1}\hat{B}_{d,1i} \tag{12}$$

where T is the sampling interval used to obtain the discrete model.

General Procedure For Exploring A Preliminary Data Set

Since process engineers may or may not be aware of the presence of integrating modes or modes behaving very closely to an integrator, the advice is to carry out a preliminary analysis of the available data to determine the presence and location of integral modes. This numerical analysis should be guided by the arguments in the Gain Matrix Estimation section and may follow the steps listed below:

1. Adopt a multivariable model order n from some a priori knowledge of the plant. The minimum advised value is $n = m.l$, which is equivalent to assuming a first-order dynamic at every transfer function $G_{ji}(s)$. Following the nomenclature in Appendix A, also adopt parameter r such that $r \geq n$ and "guess" a parameter s such that $s/r > 10$.

2. Estimate the model matrices **A**, **B**, and **C** using the SID method, then compute the eigenvalues of matrix **A**.

3. If the eigenvalues are not on the unit circle, compute the gain matrix by Equation (6); otherwise matrix **A** must be diagonalized by an appropriate transformation V, and then transformed to matrices A_d, B_d and C_d, as shown above.

4. The lines and columns of A_d, and the corresponding lines of B_d and columns of C_d, associated with the integrating modes are identified and isolated. Then, the blocks made of nonzero entries of the isolated submatrices, B_d^{int} and C_d^{int}, are used to compute the matrix $K = C_d^{\text{int}} B_d^{\text{int}}$.

5. The integrating modes can be identified in matrix **K** by looking for the nonzero entries. Furthermore, these nonzero entries, multiplied by $1/T$, provide the estimations of the integration constants K_{ji}, as indicated by Equation (12).

6. The matrices of the no-integrating part, $\tilde{A}_d$, $\tilde{B}_d$, and $\tilde{C}_d$, which are obtained by removing the lines and columns described in step 4, are used to compute the gain matrix of the system, as indicated by Equation (10).

7. Once the existence or not of integral dynamics is verified, the convenient dimension ratio s/r for the block-Hankel matrices, as well as the overall model order n, must be revised in order to obtain the desired estimation accuracy. See the application example below in the application section.

ALGORITHMS FOR MONITORING INDIVIDUAL GAIN ESTIMATIONS

In general, the task of monitoring a process system is based on estimations of the main parameters of a model obtained during earlier experiences. If the distribution density

function of an estimator is known or can be approximated, a Shewhart-type control chart can be easily set by defining a confidence interval with a given probability level. Since the gain-matrix estimator is an array of individual estimators, a rigorous procedure to conclude that a significant gain change has occurred must be carried out by individual statistical tests. These tests are typically implemented by using interval estimators or confidence intervals defined by two lateral limits around the expected value. When the estimations are repeated at time intervals, the Shewhart control chart works as an automatic hypothesis test about the posible paremeter change. In case we desire high sensitivity to detect a change, two SPC algorithms can be used: (i) the exponentially weighted moving average (EWMA) proposed by Roberts (1959), or (ii) the cumulative sum (CUSUM) algorithm proposed by Page (1954).

Multiple EWMA Control Charts

The EWMA algorithm provides a proper statistic to detect small but sustained shifts of individual gains and increases the robustness of the estimated values for those that remain unchanged. Given a sequence of estimations, $\hat{G}_{ji}(t)$ for $t = kT, (k + 1)T,...$ the algorithm executes a recursive weighted average given by the expression

$$M_{ji}(t) = \theta_{ji}\,\hat{G}_{ji}(t) + (1-\theta_{ji})M_{ji}(t-1), \qquad M_{ji}(0) = \bar{G}_{ji}, j = 1 \text{ to } l; i = 1 \text{ to } m \qquad (13)$$

where $\theta_{ji} \in [0,1]$ is a tuning parameter for handling the individual sensitivity to changes and speed of response. To compute the control limits for the statistic M_{ji}, N preliminary gain-matrix estimations must be obtained from plant data, assuming there are no gain changes during the collection period and discarding samples taken under uncommon

disturbances. Thus, every statistic variable M_{ji}, $j = 1$ to l; $i = 1$ to m, has upper and lower limits defined by

$$UL_{M_{ji}} = \bar{G}_{ji} + 3\hat{\sigma}_{M_{ji}} = \bar{G}_{ji} + 3\hat{\sigma}_{\hat{G}_{ji}}\sqrt{\frac{\theta_{ji}}{2 - \theta_{ji}}} \tag{14}$$

$$LL_{M_{ji}} = \bar{G}_{ji} - 3\hat{\sigma}_{M_{ji}} = \bar{G}_{ji} - 3\hat{\sigma}_{\hat{G}_{ji}}\sqrt{\frac{\theta_{ji}}{2 - \theta_{ji}}} \tag{15}$$

where subscripts j and i represent the output and input respectively, $\hat{\sigma}_{M_{ji}}$ is the estimated standard deviation of the statistic variable M_{ji}, and $\bar{G}_{ji}$ and $\hat{\sigma}_{\hat{G}_{ji}}$ are the estimated mean and estimated standard deviations of the estimator $\hat{G}_{ji}$ respectively. The central line for these control charts and the initial value in Equation (13) are calculated using the N preliminary gain-matrix estimations as

$$\bar{G}_{ji} = \sum_{k=1}^{N} \hat{G}_{ji}(k) / N \tag{16}$$

and the estimator standard error computed by

$$\hat{\sigma}_{\hat{G}_{ji}} = \left\{ \sum_{k=1}^{N} \left(\hat{G}_{ji}(k) - \bar{G}_{ji} \right)^2 / (N-1) \right\}^{1/2} \tag{17}$$

The "3-sigma" distance between the central line and the limits is quite an accepted, standard, and generalized convention in statistical control chart design; this distance constitutes a design parameter adopted simultaneously with the test significance level when the distribution function is known.

The monitoring procedure implies a hypothesis test every time a new value M_{ji} is obtained. If the estimator remains between limits we accept that there is no sufficient

evidence to think that a change has occurred; in contrast, one or more M_{ji} values outside

the limits must be taken as significant evidence that there has been a shift in the

population mean.

Multiple CUSUM Control Charts

Alternatively, the computational form of the CUSUM algorithm can be used for detecting

small gain changes (Montgomery, 2009). In this case, the following expressions are

applied to the sequence of gain estimations:

$$S_{ji}^{H}(t) = \max\left\{0, [\hat{G}_{ji}(t) - (\overline{G}_{ji} + b_{ji})] + S_{ji}^{H}(t-1)\right\}, \quad S_{ji}^{H}(0) = 0, \tag{18}$$

$$S_{ji}^{L}(t) = \max\left\{0, [(\overline{G}_{ji} - b_{ji}) - \hat{G}_{ji}(t)] + S_{ji}^{L}(t-1)\right\}, \quad S_{ji}^{L}(0) = 0, \tag{19}$$

where b_{ji} is used as sensitivity or tuning parameter. Typically, the design of a CUSUM

algorithm is completed with a decision variable h such that $S_{ji}^{H}(t) > h$ or $S_{ji}^{L}(t) > h$ implies

the decision that a change has occurred. This decision variable is not strictly necessary in

our applications since we focus on discovering a persistent bias between the estimator

$\hat{G}_{ji}(t)$ and the mean value $\overline{G}_{ji}$. Note from Equations (18) and (19) that while the

estimation $\hat{G}_{ji}(t)$ is inside the band $\overline{G}_{ji} \pm b_{ji}$, there will be no accumulation in $S_{ji}^{H}(t)$ or

$S_{ji}^{L}(t)$ and they will remain barely bouncing up from zero. However, as soon as the

estimation $\hat{G}_{ji}(t)$ starts going systematically to one side of the indicated band, a clear

departure from zero or upward tendency will be observed in $S_{ji}^{H}(t)$ or $S_{ji}^{L}(t)$, indicating

that a significant and persistent change has occurred.

APPLICATION TO A SIMULATED SYSTEM

MIMO System Adopted For Evaluation

The gain estimation method proposed here was tested by numerical simulation of a process plant represented by the 5×3 linear time invariant (LTI) system given in Table I. This arbitrary example is defined with more outputs than inputs because this is a frequent characteristic of actual process systems where advanced controllers, like MPCs, are applied. Furthermore, note that the representation includes important time delays and integral modes, providing in this way a challenge similar to the one encountered in real plant applications. Figure 1 shows the step responses of the transfer functions in Table I, where different but typical process dynamics are exhibited and the longest settling times can be estimated.

The convenience of using a known multivariable linear system raises the possibility of evaluating the accuracy that can be expected and the need of searching proper methodological conditions to reduce the estimation biases from "actual gain values." Hence, the numerical experience in this section is intended to highlight difficulties that might arise when using the SID method as a base to develop the desired follow-up tool, even when the application problem is free of process nonlinearities.

The availability of two sequences of data is assumed in this work. The first sequence corresponds to simulations where the system is assumed steady and without gain changes but receiving a minimal amplitude exciting signal on the manipulated variables. This first data set serves for designing the detection algorithms and the control limits following the

statements in the Algorithms for Monitoring Individual Gain Estimations section above. The second data set comes from simulations where three gains are arbitrarily changed in order to test the sensitivity of the proposed methods to detect unexpected process gain shifts.

Data Collection Setup

The monitoring strategy analyzed here assumes the possibility of having operation periods where a persistent excitation can be injected to the system. Hence, a sequence of batch-wise data collections must be available as relevant information for applying the multiple gains follow-up procedure.

The setup for data collection assumed in this work is shown in Figure 2. Besides, and most important, a modified pseudo-random binary signal (MPRBS) is imposed on the manipulated variables to obtain output responses with enough variability to expose cause-effect reactions. The modification introduced to the standard PRBS and explained in detail in the next subsection, is motivated by the possibility of finding integral mode dynamics frequently present in chemical processes, which could drive the open-loop system unstably if they are not properly compensated for. Note it is assumed here that there is no previous knowledge about the presence of integral dynamics imbedded in the data sets.

Input Test Signals

As in many identification procedures, the experience starts by injecting exciting signals to the input variables and registering both input and output variables, as shown in Figure 2. In order to come close to typical signals obtained from actual process systems, Gaussian noise is arbitrarily added to the output variables. The exciting signals used in this application are similar to the standard PRBS but conditioned to comply with two requirements: (i) to be an independent and persistent excitation for each input, and (ii) to be built to avoid long time intervals during which the plant integral modes might take the system outside the allowed operating region.

The first requirement for the input signal $u(k)$ is associated with the linear independence of the rows in the block Hankel matrix of the input data (see Appendix A). If this condition does not hold, then matrix $U_f U_f^T$, whose inverse is necessary to obtain the orthogonal and oblique projections, becomes singular and the estimation problem does not provide a unique solution. To avoid this difficulty, the input is required to be "rich enough," which mathematically means that the persistent excitation condition defined in Appendix C is fulfilled.

The second requirement is solved by creating zero-integral cycles, built by a signal with opposite-sign parts such that, every time a cycle is completed, the integral of the signal goes to zero. The plot of u_1 in Figure 3 shows this particular feature by including vertical dashed lines to visualize the compensated cycles. The random characteristics of the signal appear with the length and the direction of the first movement in every cycle. These modified PRBS signals (MPRBS) start with different seeds for every input to avoid

simultaneous similar sequences, escaping in this way having correlations between the input signals. The frequency band is determined by taking the upper and lower bounds associated to the shortest and longest settling times respectively. Regarding the signal amplitude, the added excitation is such that the signal-to-noise ratio (SNR) in each output variable is large enough to distinguish the deterministic signal from noise. The adopted signal amplitude is such that it produces SNRs going from 20 to 50 dB for the different outputs (Davis and Davis, 1997). Figure 3 shows the MPRBS signals previous to being affected by the individual amplitude factors.

Model Order and Hankel Matrix Rectangularity

Numerical analysis is recommended for determining the convenient dimensions for the block Hankel matrices required by the SID method, as well as the overall model order. These dimensions are discussed in detail in Appendix A, where r and s are conditioned to $r \geq n$ and $s \gg r$ <AQ: please check symbol that was not translated between s and r>, n being the adopted order of the system model. Recall that the interest focuses here on estimating the process gains and that a detailed identification of rapid process modes is not necessary. Consequently, some numerical experiments are suggested for finding the lower model order n and s/r ratio that provides acceptable estimation accuracy. Note that s/r defines the "rectangularity" of the Hankel matrices; the greater the ratio, the more rectangular the matrix.

Though there are alternative methods, a simple approach to quantify the accuracy of the estimated gain matrix consists in averaging the individual absolute errors of the gains estimations, i.e.,

$$AAE = \frac{1}{l.m} \sum_{j=1}^{l} \sum_{i=1}^{m} \left| \hat{G}_{ji} - G_{ji} \right| \qquad (20)$$

where G_{ji} is the true gain value between the output j and the input i.

Figure 4 shows how the function AAE varies with n and s/r when the SID method is used for the estimation of the gains of the system in Table I. This result is obtained for a 2 min sampling interval, which could be close to the typical practice when large processes of standard complexity are considered. This time interval means that the data set available for one gain-matrix estimation per day can be composed by at least 600–700 readings per variable, though some of them might have to be excluded due to instrument failures, process upsets, and other issues.

Notice from Figure 4 that an insignificant variation of the average gain error is observed when both the model order n and the ratio s/r go above 15 or 16; from there on, an absolute error below 0.035 is obtained as average over the 15 gain estimations. The fact that the accuracy is quite good even for $n = 10$ when the ratio s/r is large enough implicitly confirms that not all the system modes need to be considered for good gain estimations. This preliminary analysis produces a considerable reduction of the computational load by helping to avoid unnecessarily large parameters n and s/r and contributes to implementation in large process systems.

The absolute errors used for drawing the function in Figure 4 were computed by using, as true gain values, those found in the transfer functions of Table I. However, in a real world problem the true gain values will never be known, only estimated. Hence, the numerical analysis to determine the convenient *s/r* and *n* pair has to be done without knowing the true gain values. Fortunately, the asymptotic behavior observed in the *AAE* function, as *s/r* and *n* increases, is maintained independently of the "true values" used as references to compute the function. This means that any difference between theses reference values and the actual gains simply shift the function up in Figure 4 without changing the shape significantly, thus retaining the capacity for helping the selection of the values *s* and *n*. Since the main objective is the detection of changes in the process gains that move away from those in use by the MPC controller, the use of the MPC gains as references to compute the *AAE* function becomes a natural adoption when applying this procedure to a real process under MPC control.

Individual Gain Estimations During Undisturbed Operation

Notice in the system of Table I that the transfer functions $G_{13}(s)$ and $G_{32}(s)$ are both pure integrators with integration constants 1.0 and 2.0 respectively. Once they are localized and estimated, the remaining task is the estimation of the model gains associated with the stable modes as described by the analysis presented above.

Figure 5 shows the estimations performed "once a day" with data taken from the first sequence of simulations, where the process is assumed steady and without gain changes. From a simple inspection it is possible to observe high accuracy in most cases.

Results Obtained with EWMA

As indicated previously, each gain estimator is a statistical variable trying to determine the actual gain value. Consequently, a reliable statement about the change of a specific gain G_{ji} must be made by hypothesis testing on the mean of the estimator $\hat{G}_{ji}(t)$. The sensitivity of this statistical inference can be significantly improved when the test is performed through the EWMA algorithm as soon as a new estimation is available. In this numerical experience a few gain changes were introduced arbitrarily in the "process plant" represented by the transfer functions in Table I; these changes are as follows: (1) a positive ramp change of 3% was simulated in gain G_{12} from day 5 to 8, (2) a positive ramp change of 3% was implemented for gain G_{41} from day 10 to 13, and finally, (3) a negative ramp change of 5% was made to gain G_{53} from day 5 to 8. Figure 6 shows the 15 gain estimations per day for 25 days; the plots clearly show how the EWMAs corresponding to the disturbed gains go outside their control limits. In general, the scalar EWMA works well in practice when $0.05 \leq \theta_{ji} \leq 0.3$; in this application the observed sensitivity was obtained by setting θ_{ji} to 0.3 for all the cases.

Results Obtained With CUSUM

The same numerical test was repeated but using the individual CUSUM. The results of applying CUSUM are shown in Figure 7, where both $S_{ji}^{H}(t)$ and $S_{ji}^{L}(t)$ are plotted for the 25 days assumed by the overall simulation time. To obtain good sensitivity, the values of b_{ji} were set to 0.05. Though no decision limit h was used in these plots, as in traditional

applications, it is apparent that the detection and localization objectives were achieved (see the plots corresponding to G_{12}, G_{41}, and G_{53}). In addition, if the changes are permanent, the importance of amplitude of the gain displacement can be quantified from the slope presented by the detection signal (Lucas, 1982).

CONCLUSIONS AND FUTURE WORK

The results obtained in this work provide evidence that the stationary gains of a multivariable process can be successfully estimated by using the subspace identification method with a batch-wise data-gathering mode. Since chemical processes frequently include capacities or accumulators that generate dynamic behaviors very closely described by integral modes, the procedure for estimating the gain-matrix includes cases where there is no previous knowledge about the presence of integral dynamics imbedded in the data sets. Therefore, specific attention is given to the treatment required by the identification when there are dynamic evidences of integral-mode behavior. A new approach based on subspace identification was developed to overcome this difficulty, which demonstrates that the identification task is possible under open-loop operation yielding reliable estimations.

The numerical experience also shows that univariate-sensitive techniques like the statistical control charts of EWMA or CUSUM applied to every element of the estimated gain matrix can be considered a promising method for detecting and localizing process gain changes. These results raise the possibility of developing monitoring strategies capable of detecting changes of the most important parameters of the model used by LP-

MPC, particularly in the LP module when determining the control targets. The methodology developed here for online monitoring of process gains shows hopeful results, and the extension to closed-loop identification together with the explicit consideration of constraints is a challenge to be analyzed and developed in the future.

ACKNOWLEDGMENTS

APPENDIX A: N4SID NOTATION

The notation used in this article follows that commonly used in the extensive literature available about subspace identification methods (Van Overschee and De Moor, 1996). For instance, the block Hankel matrix of a single-input signal is defined and written as

$$U_{0|2r-1} \triangleq \begin{pmatrix} u_0 & u_1 & u_2 & ... & u_{s-1} \\ u_1 & u_2 & u_3 & ... & u_s \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{r-1} & u_r & u_{r+1} & ... & u_{r+s-2} \\ u_r & u_{r+1} & u_{r+2} & ... & u_{r+s-1} \\ u_{r+1} & u_{r+2} & u_{r+3} & ... & u_{r+s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{2r-1} & u_{2r} & u_{2r+1} & ... & u_{2r+s-2} \end{pmatrix} = \begin{pmatrix} U_{0|r-1} \\ \hline U_{r|2r-1} \end{pmatrix} = \begin{pmatrix} U_p \\ \hline U_f \end{pmatrix} = \begin{pmatrix} U_{0|r} \\ \hline U_{r+1|2r-1} \end{pmatrix} = \begin{pmatrix} U_p^+ \\ \hline U_f^- \end{pmatrix} \qquad (A.1)$$

where the indexes r and s are such that $r \geq n$ and $s \gg n$, and n is the assumed order of the system. Note that indexes p and f stand for "past" and "future" respectively; this is because each column of matrix U_p is composed of r elements previous to the following r elements being part of U_f. Notice also that the partition can be made one line below, and this be indicated as shown by the last two equalities on the right. This matrix structure

was also extended to the block Hankel matrix for a single-output variable, in this case

using the notation $Y_{0|2r-1}, Y_p, Y_f, \ Y_p^{+}$, and Y_p^{-}. The above notation serves also to define the

block Hankel matrix of past input and output data, i.e.,

$$W_p \triangleq \begin{pmatrix} U_p \\ Y_p \end{pmatrix} \tag{A.2}$$

In a similar way, the state sequence

$$X_r = \begin{pmatrix} x_r & x_{r+1} & \cdots & x_{r+s-2} & x_{r+s-1} \end{pmatrix} \in \mathbb{R}^{n \times s} \tag{A.3}$$

is partitioned by setting $X_p = X_0$ and $X_f = X_r$.

The subspace identification algorithm needs other important matrices: the observability

matrix $\Gamma_r \in \mathbb{R}^{lr \times n}$,

$$\Gamma_r = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{r-1} \end{pmatrix} \tag{A.4}$$

and two low-triangular block Toeplitz, $H_r^d \in \mathbb{R}^{lr \times mr}$ and $H_r^s \in \mathbb{R}^{lr \times mr}$, which are written as

follows:

$$H_r^d = \begin{pmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ CA^{r-2}B & CA^{r-3}B & CA^{r-4}B & \cdots & D \end{pmatrix} \tag{A.5}$$

and

$$H_r^s = \begin{pmatrix} I & 0 & 0 & \cdots & 0 \\ CK & I & 0 & \cdots & 0 \\ CAK & CK & I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ CA^{r-2}K & CA^{r-3}K & CA^{r-4}K & \cdots & I \end{pmatrix} \qquad (A.6)$$

Assuming the pair $\{A,C\}$ is observable and $r \geq n$, then Γ_r is a full column-range matrix, i.e., $rank(\Gamma_r) = n$. We need still to explain the notation A/B, which stands for the orthogonal projection of the row space of A into the row space of B:

$$A/B \triangleq A\Pi_B = AB^T \left(BB^T\right)^{\dagger} B \qquad (A.7)$$

where $(\bullet)^{\dagger}$ denotes pseudo-inverse of the matrix $(\bullet)$.

further, $A/_C B$ denotes an oblique projection, i.e., the projection of the row space of A along the row space of B on the row space of C. This can be defined as:

$$A/_B C \triangleq \left[A/B^{\perp}\right]\left[C/B^{\perp}\right]^{\dagger} C \qquad (A.8)$$

where $B^{\perp}$ represents the orthogonal complement of matrix B. See Van Overschee and De Moor (1996) for further details.

APPENDIX B: MODEL MATRIX ESTIMATIONS

This appendix highlights the main concepts of the subspace identification method used in this article. Please note the convenience of being familiar with the nomenclature explained in Appendix A in going through this appendix. Let us start by indicating that an iterative substitution of Equations (1) and (2) allows writing the expression

$$Y_f = \Gamma_r X_f + H_r^d U_f + E_f$$

(B.1)

where the subscript f denotes "future horizon." The matrices $Y_f \in \mathbb{R}^{lr \times j}$, $U_f \in \mathbb{R}^{mr \times s}$, and

$E_f \in \mathbb{R}^{lr \times s}$ are the output, input, and noise block Hankel matrices respectively. H_r^d is a

low-triangular Toeplitz matrix constructed from the system impulse responses.

The information about the system is mainly in the first term on the right of expression

(B.1), which includes the extended observability matrix, Γ_r, and the state sequence, X_f.

Following the nomenclature in Appendix A, the oblique projection of the row space of

Y_f along the row space of U_f on the row space of W_p can be written as follows:

$$Y_f /_{U_f} W_p = \Gamma_r X_f /_{U_f} W_p + H_r^d U_f /_{U_f} W_p + E_f /_{U_f} W_p \tag{B.2}$$

If the noise E_f is assumed independent from the past input U_p, the past output Y_p, and

the future input U_f, and by the property of the oblique projection $U_f /_{U_f} W_p = 0$, the last

two terms are null, then the following relationship can be obtained:

$$Y_f /_{U_f} W_p = \Gamma_r X_f /_{U_f} W_p = \Gamma_r \hat{X}_f \tag{B.3}$$

This result means that the space column of Γ_r is the same as the space column of

$Y_f /_{U_f} W_p$, which can be estimated from input-output data. Then, applying singular value

decomposition, the left-hand side can be written as

$$Y_f /_{U_f} W_p = U S V^T = U S^{1/2} S^{1/2} V^T \tag{B.4}$$

from which the extended observability matrix Γ_r can be estimated by

$$\hat{\Gamma}_r = U S^{1/2} \tag{B.5}$$

Once this matrix is obtained, the model matrices in (1) and (2) can be estimated in the following order: C, A, B, and D. The matrix C is obtained directly from the first row block of Γ_r, i.e.,

$$C = \Gamma_r(0:l-1, 0:n-1) \tag{B.6}$$

The matrix A is then obtained using the invariant displacement property of Γ_r:

$$\overline{\Gamma}_r = \underline{\Gamma}_r A \tag{B.7}$$

where $\overline{\Gamma}_r \in \mathbb{R}^{l(r-1)\times n}$ and $\underline{\Gamma}_r \in \mathbb{R}^{l(r-1)\times n}$ are matrices obtained from the last $l(r-1)$ and the first $l(r-1)$ row blocks of Γ_r respectively. Hence, matrix A can be estimated by the mean squared solution as follows:

$$A = \underline{\Gamma}_r^{\dagger} \overline{\Gamma}_r \tag{B.8}$$

where $\underline{\Gamma}_r^{\dagger} = \left(\underline{\Gamma}_r' \underline{\Gamma}_r\right)^{-1} \underline{\Gamma}_r'$ is a pseudo-inverse matrix.

Now, left and right multiplying Equation (B.1) by $\Gamma_r^{\perp}$ and $U_f^{\dagger}$ respectively gives

$$\Gamma_r^{\perp} Y_f U_f^{\dagger} = \Gamma_r^{\perp} \Gamma_r X_f U_f^{\dagger} + \Gamma_r^{\perp} H_r^d U_f U_f^{\dagger} + \Gamma_r^{\perp} E_f U_f^{\dagger} \tag{B.9}$$

Notice that $\Gamma_r^{\perp} \in \mathbb{R}^{(lr-n)\times lr}$ is a complete range matrix satisfying $\Gamma_r^{\perp} \Gamma_r = 0$ and $U_f U_f^{\dagger} = I$. Hence, assuming negligible noise, the above expression simplifies to

$$\underbrace{\underbrace{\Gamma_r^{\perp} Y_f U_f^{\dagger}}_{\in \mathbb{R}^{(lr-n)\times mr}}}^{Data} = \underbrace{\Gamma_r^{\perp}}^{Data} H_r^d \tag{B.10}$$

This is an over determinate system of linear equations where matrices B and D are the unknowns, which can be rewritten as follows:

$$\left(M_1\ M_2 \ldots M_r \right) = \left(L_1\ L_2 \ldots L_r \right) \begin{pmatrix} D & 0 & \cdots & 0 \\ CB & D & \cdots & 0 \\ CAB & CB & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ CA^{r-2}B & CA^{r-3}B & \cdots & D \end{pmatrix}$$

and rearranging

$$\underbrace{\begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_r \end{pmatrix}}_{\in \mathbb{R}^{r(lr-n) \times m}} = \underbrace{\begin{pmatrix} L_1 & L_2 & \cdots & L_{r-1} & L_r \\ L_2 & L_3 & \cdots & L_r & 0 \\ L_3 & L_4 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ L_r & 0 & 0 & \cdots & 0 \end{pmatrix}}_{\in \mathbb{R}^{i(li-n) \times li}} \underbrace{\begin{pmatrix} I_l & 0 \\ 0 & \Gamma_r \end{pmatrix} \begin{pmatrix} D \\ B \end{pmatrix}}_{\in \mathbb{R}^{lr \times (l+n)}}, \tag{B.11}$$

where $M_k \in \mathbb{R}^{(lr-n) \times m}$ and $L_k \in \mathbb{R}^{(lr-n) \times l}$. This equation system can then be solved for B and D using a least squares approach.

APPENDIX C: PERSISTENT EXCITATION

The persistent excitation condition necessary to perform identification can be (mathematically) defined as:

Definition:

The input sequence $\{u(k)\}$, $k = 1, 2, 3, \ldots$, is a persistent excitation of order n if:

1)   The following limits exist:

$$r_u(\tau) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} u(k+\tau) u^T(k)$$

(C.1)

where $r_u(\tau)$ is the auto regression function for the time displacement $\tau = 0, 1, \cdots, n-1$, and

2)   The auto regression matrix

$$
R_u(n) = \begin{bmatrix}
r_u(0) & r_u(1) & \cdots & r_u(n-1) \\
r_u(-1) & r_u(0) & \cdots & r_u(n-2) \\
\vdots & & \ddots & \\
r_u(1-n) & r_u(2-n) & \cdots & r_u(0)
\end{bmatrix} \tag{C.2}
$$

is positive definite. In the subspace identification context, the persistent excitation is

directly related to the rank of the Hankel matrices involved in the computation.

REFERENCES

Darby, M. L., and Nikolaou, M. (2012). MPC: Current practice and challenges, *Control Eng. Pract.*, **20**, 328–342.

Davis, D., and Davis, C. (1997). *Sound System Engineering*, 2nd ed., Focal Press, Boston.

Favoreel, W., De Moor, B., and Van Overschee, P. (2000). Subspace state space system identification for industrial processes, *J. Process Control*, **10**, 149–155.

González, A. H., Odloak, D., and Marchetti, J. L. (2006). Predictive control applied to heat-exchanger networks, *Chem. Eng. Process*., **46**, 661–671.

Katayama, T. (2005). *Subspace Methods for System Identification*, Springer, Berlin.

Larimore, W. E. (1990). In *Identification, Filtering and Adaptive Control, Proceedings of the 29th IEEE Conference on Decision and Control*, 596–604, Institute of Electrical and Electronic Engineers, Piscataway, N.J.

Lucas, J. M. (1982). Combined Shewhart-CUSUM quality control scheme, *J. Qual. Technol.*, **14**, 51–59.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*, 6th ed., John Wiley & Sons, Hoboken, N.J.

Nikandrov, A., and Swartz, C. L. E. (2009). Sensitivity analysis of LP-MPC cascade control systems, *J. Process Control*, **19**, 16–24.

Page, E. S. (1954). Continuous inspection schemes, *Biometrika*, **41**, 100–114.

Qin, S. J., and Badgwell, T. A. (2003). A survey of industrial model predictive control technology, *Control Eng. Pract.*, **11**, 733–764.

Roberts, S. W. (1959). Control charts based on geometric moving averages, *Technometrics*, **1**, 239–250.

Van Overschee, P., and De Moor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, *Automatica*, **30**(1), 75–93.

Van Overschee, P., and De Moor, B. (1996). *Subspace Identification for Linear Systems: Theory, Implementation, Applications*, Kluwer Academic Publishers, Boston.

Verhaegen, M., and Dewilde, P. (1992). Subspace model identification, Part I: The output-error state-space model identification class of algorithms, *Int. J. Control*, **56**, 1187–1210.

Ying, C. M., and Joseph, B. (1999). Performance and stability analysis of LP-MPC and QPMPC cascade control systems, *AIChE J.*, **45**, 1521–1534.

Table I. Multivariable linear system used for testing the method for detecting gain

changes

$\dfrac{-0.8}{15s^2+3s+1}$	$\dfrac{4.7}{10s^2+7.8s+1}$	$\dfrac{1}{s}$
$\dfrac{1.4e^{-10s}}{8s^2+5s+1}$	$\dfrac{0.8e^{-10s}}{10s^2+3s+1}$	0.0
$\dfrac{1.5}{7.5s+1}$	$\dfrac{2}{s}$	$\dfrac{0.9}{15s^2+10s+1}$
$\dfrac{4.0}{7.8s+1}$	0.0	$\dfrac{2.5}{5s+1}$
$\dfrac{-0.3}{10s^2+4s+1}$	$\dfrac{5.0e^{-5s}}{16s^2+5s+1}$	$\dfrac{-2.8}{9.8s^2+10s+1}$

Figure 1. Step responses corresponding to the transfer functions in Table I.
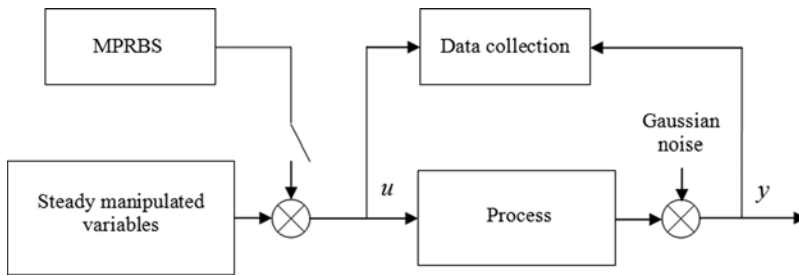
Figure 2. Open-loop batch-wise data collection.
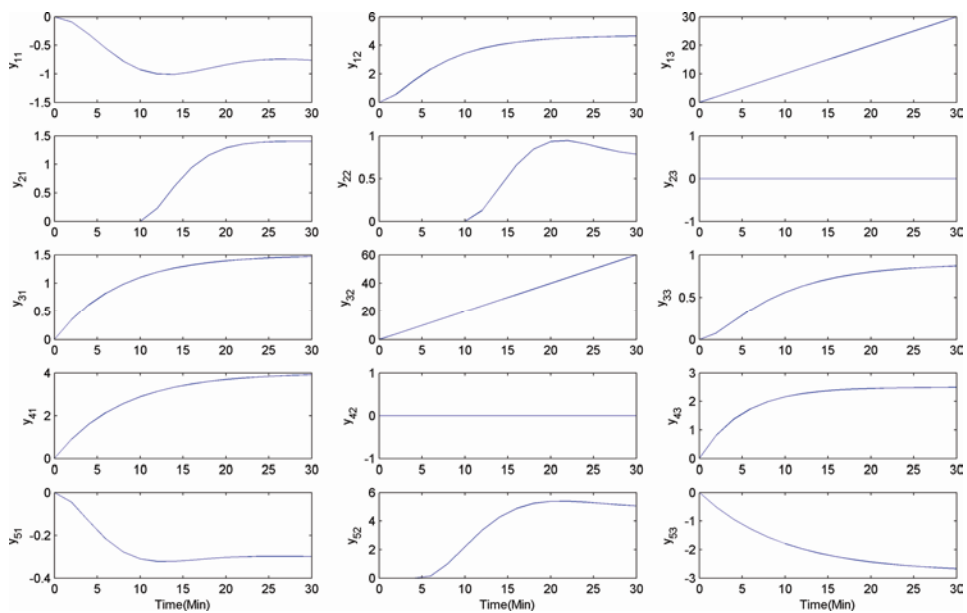
Figure 3. Modified PRBS signals injected to the system in Table I.
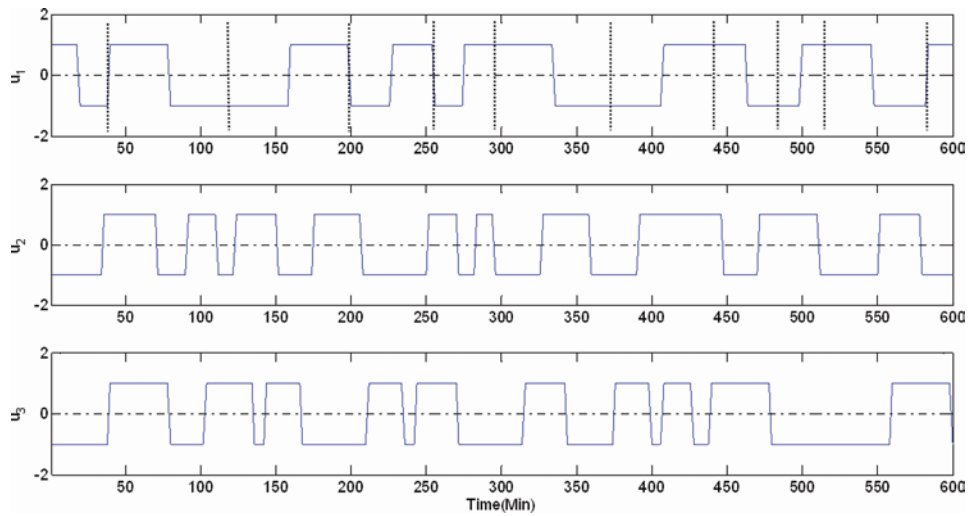
Figure 4. Average absolute error of the gain estimations for different values of the model
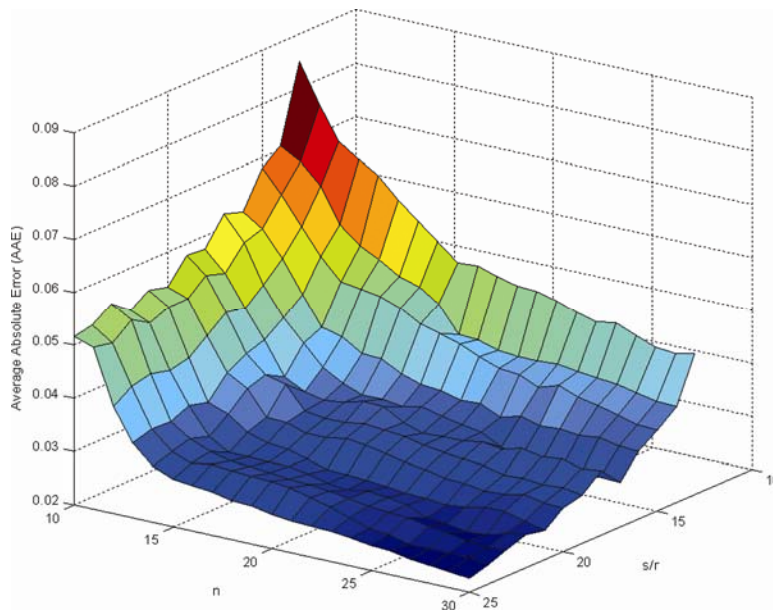
order *n* and the ratio *s/r*.

Figure 5. Preliminary sequence of gain estimations. The dashed lines indicate true values.
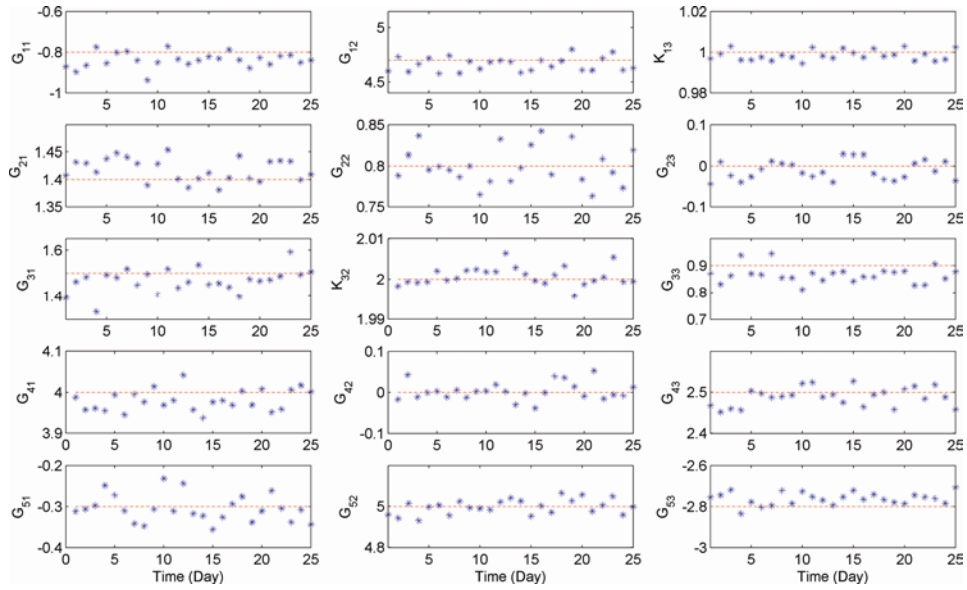
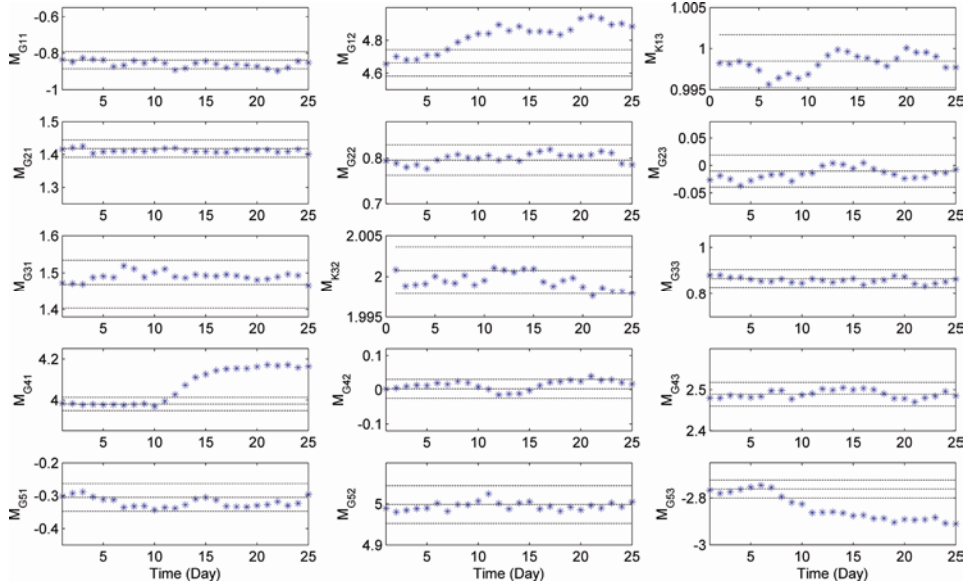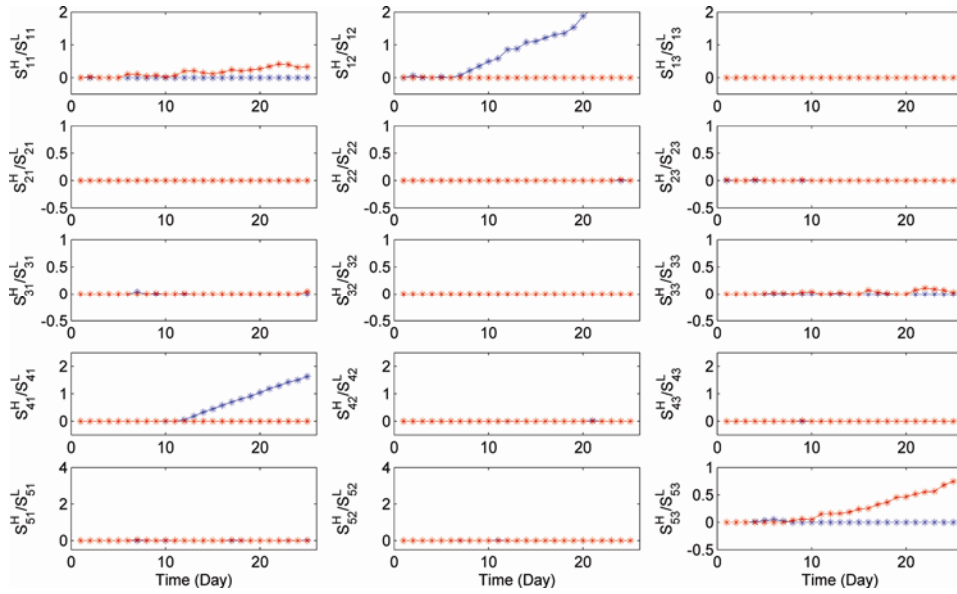Figure 6. EWMA charts showing the response to ramp changes in G_{12}, G_{41}, and G_{53}.

Figure 7. CUSUM charts showing the response to ramp changes in G_{12}, G_{41}, and G_{53}.