Extraction of informative data subsets for use in data-driven control

Cristiane Silva Garcia[1] and Alexandre Sanfelice Bazanella[1]

Abstract— **Performing a specific experiment to collect data to estimate the controller's parameters can be a difficult and, sometimes, an undesired task. Instead of that, an attractive idea is to use data gathered from normal operation routines. In some cases these data may have enough information to estimate these parameters. Therefore, the goal of this work is to apply the metrics already existent in the system identification literature to the controller's estimation problem and to propose a new metric, to search for informative subsets of data from the entire collected data set. In the present work, the parameters were estimated using the virtual reference feedback tuning (VRFT) method. In order to attest the feasibility of the proposed solution some simulation case studies are also presented.**

I. INTRODUCTION

In general, within the optimal data-driven control framework, the task of estimating the controller's parameters requires the execution of a specific experiment in a plant to collect data. This experiment requires a sufficient rich signal to be applied to the system, which, in most cases, differs from the signal that is applied to the system in normal operating mode. In several cases, this is a very costly task and, sometimes, may be even impossible because of some reasons. For example, interrupting the normal operation to perform the specific experiment can be a costly task.

Therefore, a better option is to use data collected from normal operation instead. In industrial processes, data gathered from normal operation routines are usually stored in a data base, being already available for free. Often, these data provide relevant information that can be used to identify the parameters of a controller.

The problem of searching for informative subsets within the entire data set has already been treated within the system identification framework. In [1] the authors proposed a data removal technique that is used to discard the data that are strongly dependent of the noise. The technique uses singular value decomposition (SVD) to extract the singular values of a regressor matrix. Then the slope of the smallest singular value as a function of time is used as a metric to remove the data that does not have relevant information. It was shown that using all the data may worsen the quality of the obtained parameters estimation because it would increase the bias of the estimated parameters.

In [2] an algorithm was developed to find relevant intervals of data to system identification within a historical data base. The algorithm searches for variations on the input and output signals, and also uses the condition number of the information matrix to determine if a sequence of data is informative enough. Besides, combined to the two metrics mentioned above, the algorithm verifies how much the input and output signals are correlated, and uses it as a metric to define the previously selected data sequence as useful, and also as a quality indicator for this sequence. An extension of the work in [2] was presented in [3], where a forgetting factor was added to the estimation of the Laguerre model and a noise model was introduced.

Within the work presented in [4] the algorithm proposed in [2] was used to segment the data, aiming to detect when the process model changed, and to identify the different models for the plant corresponding to the segments found. In [5] the authors used the condition number of the Fischer information matrix to determine, from the amount of data collected from normal operating routines, the sequences of data that are relevant to identify the system model. The same work also suggested a value for the threshold of the condition number. In the work developed in [6] a new method to search for informative data addressed to system identification was presented. In that work the reciprocal condition number is also used as a metric to determine the informative subset of data. In that case, only a few input changes are considered, and it was shown that the bias of the estimated parameters decreased using the selected subset. In [7] an extension of [6] to the multiple-input multiple-output (MIMO) case is presented. In [8] a criterion to search for informative subsets in data gathered from normal operation routines is presented. This criterion is based on finding significant magnitude changes in the input and output signals. In the work developed in [9] a rank test was presented to delimit the informative data subset. In that case, the system is identified using subspace system identification.

Searching for informative intervals of data is a task that has not received much attention in the data driven control framework. Besides that, it would be good to define some requirements, as for example, the amount of data to be collected and used in the parameters estimation. Moreover, it is useful to know if the amount of data used for estimation of the parameters can interfere in the quality of the obtained estimation.

With all that in mind, the present work proposes the employment of the smallest singular value and the reciprocal condition number of the information matrix to the controller's parameters estimation problem. This was inspired by the works developed in [1] and [3], which addressed the

[1]The authors are with the department of Automation and Energy – Federal University of Rio Grande do Sul. Av. Osvaldo Aranha 103 – CEP:90035-190 – Porto Alegre, RS – Brazil {cristiane.garcia, bazanella}@ufrgs.br

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

system identification problem. Besides that, a new metric is suggested in this work, based also in the reciprocal condition number.

An advantage of using the data-driven approach, as the name suggests, is that the controller is estimated directly from the data. Thus, in the case of using a simple controller class as proportional-integral (PI) or a proportional-integral-derivative (PID) reduces the complexity of the information matrix compared to the system identification approach where, in general, a high order process is modeled.

In this paper, the parameters are calculated using the Virtual Reference Feedback Tuning (VRFT) method, so the information matrix and the regressor vector used are generated by this method. The VRFT is a non-iterative data-driven method, which means that, in the ideal case, the data of a single experiment is enough to calculate the controller's parameters [10]. This way, it is suitable to be used with normal operation data.

Therefore, the metrics mentioned above are used to delimit the informative amount of data to be used to identify the controller's parameters. Moreover, the value of the cost function is used as a measure to investigate the effect of reducing the number of samples on the estimation of the parameters. Some simulation results are presented for different scenarios in order to demonstrate the feasibility of the proposed solution.

This paper is organized as follows: section II presents a briefly review of the methodologies proposed and presents the VRFT method. In section III it is explained how the metrics are applied to the addressed problem. The results of some simulation experiments are presented in section IV. Finally, the conclusion and the future work are discussed in section V.

II. PRELIMINARIES

A. Smallest singular value approach

The aim of this section is to provide a brief explanation of the data discarding criterion, addressed to the system identification framework, developed in [1]. In that work, all the theoretical formulation was designed considering that: the system is single-input single-output (SISO) and stable, all data is from open loop simulations, and the experiment input comprises only a few step changes and long dwell times.

The system to be modeled is an auto-regressive with exogenous input (ARX), and its output y and output predictor $\hat{y}(\theta)$ are given in vector form as

$$y = \Phi\theta_0 + \epsilon \tag{1}$$

$$\hat{y}(\theta) = \Phi\theta \tag{2}$$

where, $y = [y(1), \ldots, y(N)]^{\mathrm{T}}$ is the system output vector, $\theta_0 \in \mathcal{R}^{n_p}$ is the true parameters vector, $\epsilon = [\epsilon(1), \ldots, \epsilon(N)]^{\mathrm{T}}$ is the model error, $\hat{y}(\theta)$ is the predicted output vector for any parameter vector θ, and $\Phi = [\phi(1), \ldots, \phi(N)]^{\mathrm{T}} \in \mathcal{R}^{N\times n_p}$ is the regressor matrix, where the regressor vector is given by

$$\phi(t) = \left[-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b)\right]^{\mathrm{T}}$$

The parameters are estimated through the standard least squares prediction error criterion given by

$$\hat{\theta}(N) \triangleq \arg\min_{\theta\in\mathcal{R}^{n_p}} J(\theta, N) = \frac{\left\|\varepsilon(\theta)\right\|_2^2}{2N} = \frac{\left\|y - \hat{y}(\theta)\right\|_2^2}{2N}$$

where $\varepsilon(\theta)$ is the prediction error in vector form and $\|\cdot\|_2$ is the Euclidean norm.

The solution for this problem can be written using the pseudo-inverse $\Phi^{\dagger}$ of the regressor matrix Φ as $\hat{\theta} = \Phi^{\dagger}y$. Therefore, applying the singular value decomposition technique in Φ one can get

$$\Phi = U\Sigma V^{\mathrm{T}}$$

where, $\Sigma \in \mathcal{R}^{n_p\times n_p}$ is the singular value matrix formed by $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{n_p})$, and $\sigma_i^2 = \lambda_i(\Phi^{\mathrm{T}}\Phi) > 0$ for $i = 1, \ldots, n_p$. The matrices U and V are orthogonal matrices, U is the left singular matrix of Φ and V is the right singular matrix of Φ. The SVD technique was used to make a similarity transformation, that is, to lead the $\hat{\theta}$ parameters vector to the eigenparameters space through $\hat{\theta}_V = V^{\mathrm{T}}\theta$. Using the eigenparameters representation and based on the excitation assumption that $\sigma^2 \ll \sigma_i^2,\ \ i = 1, \ldots, n_p$, that is, each eigensubspace energy σ_i^2 is much larger than the system noise power σ^2, it was shown that:

- Each eigensubspace energy σ_i^2 is composed by a term due to the input excitation in that eigensubspace and by a term proportional to the noise energy and the number of samples. This way, if the increase of σ_i^2 is small it may be because it is affected only by the noise energy term, and because this term only grows with time it can deteriorate the parameters estimation.
- The mean of the eigenparameters $\hat{\theta}_{Vi}$, where i represents the i-th component of $\hat{\theta}_V$, are nearly independent of each other. Although, some correlation between them exists due to the unmodeled noise.
- The bias of $\hat{\theta}_{Vi}$ may increase with the number of samples used if the input energy on that eigensubspace is not significant, as will be illustrated, for the controller's parameters estimation problem, in section III.
- In general, the variance of $\hat{\theta}_{Vi}$ decreases as the number of samples is increased.
- When there is a singular value that is significantly smaller then the others $\sigma_{i_{min}} \ll \sigma_i$, $i_{min} \neq i$, the eigenparameter $\hat{\theta}_{Vi_{min}}$ is the most affected by bias and variance.

As $\hat{\theta}_{Vi_{min}}$ is the most poorly estimated eigenparameter and it influences in the accuracy of the actual parameters, once the eigenparameters $\hat{\theta}_V$ lead to the actual parameters $\hat{\theta}$ through an orthogonal transformation, it is reasonable to only consider the time variations of $\sigma_{i_{min}}$. This way, the discarding criterion is defined as

$$\underline{\sigma}^2(N) - \underline{\sigma}^2(N-1) < \eta_c \tag{3}$$

where, $\underline{\sigma}$ is the smallest singular value of the regressor matrix Φ, N is the number of samples and η_c is a suitable threshold. Therefore, the regressor is discarded if the inequality (3) is

satisfied. The value of η_c may be chosen as a few orders of magnitude larger than the smallest slope of the graph of $\underline{\sigma}^2(N)$.

B. Condition number approach

The second approach is presented in [3] and a brief explanation of the method is given bellow. In this case, the following assumptions are made: the system is SISO, the system can be well described with a linear model, and it is assumed that the input signal is driven by a sequence of steps with dwell times.

In that work, the parameters vector is estimated through the recursive least squares (RLS) method given by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \bar{R}^{-1}(t)\phi(t)\varepsilon(t,\hat{\theta}_{t-1})$$
$$\bar{R}(t) = \lambda\bar{R}(t-1) + \phi(t)\phi^{\mathrm{T}}(t)$$

where, $\varepsilon(t,\hat{\theta}_{t-1})$ is the prediction error, $\phi(t)$ is the regressor vector, $\bar{R}(t)$ is the information matrix, and λ is a weighting factor with $0 < \lambda < 1$. In a simple way, the algorithm can be defined through the following steps:

Search for any input step change: First, look for steps in the input signal. This way, each identified change in the input and output signals generates a data sequence. A search for an informative subset occurs within each data sequence, which is the next step.

Condition number test: In this test, the reciprocal condition number of the information matrix $\bar{R}(t)$ is calculated increasing the number of samples and it is tested through a threshold as

$$\gamma(t) = \frac{\underline{\sigma}(t)}{\overline{\sigma}(t)} > \eta_n \tag{4}$$

where $\underline{\sigma}$ is the smallest singular value, $\overline{\sigma}$ is the biggest singular value of $\bar{R}(t)$, and η_n is an appropriate threshold. Each data subset comprises one sequence truncated as soon as the above condition stops being true.

C. VRFT method

The VRFT is a non-iterative data driven method, that is, in the ideal case the parameters can be estimated using the data collected from a single experiment. This way, no special experiment is required. It can be applied to routine operation data provided that the following considerations are made: the controller is linearly parametrized, the ideal controller belongs to the controller class, and the system is not affected by noise [10].

Thus, the linearly parametrized controller can be described as $C(q,\rho) = \rho^{\mathrm{T}}\bar{C}(q)$ where, ρ is the parameters vector, $\bar{C}(q)$ is a vector representing the controller class, and $C(q,\rho)$ is the controller transfer function.

The method's goal is to solve the following problem

$$\min_{\rho} J_y(\rho) \triangleq \bar{E}\left[y(t,\rho) - y_d(t)\right]^2 \tag{5}$$

where $J_y(\rho)$ is the reference performance criterion to be minimized, $y(t,\rho) = C(q,\rho)r(t)$ is the closed loop process output signal, and $y_d(t) = T_d(q)r(t)$ is the desired output, where $T_d(q)$ is the reference model.

The VRFT method transforms the problem of minimizing a cost function $J_y(\rho)$ into a least squares (LS) identification of the controller $C(q,\rho)$, which consists in minimize $J^{\mathrm{VR}}(\rho)$. The cost function $J^{\mathrm{VR}}(\rho)$ is defined as

$$J^{\mathrm{VR}}(\rho) = \bar{E}\left[u(t) - \rho^{\mathrm{T}}\varphi(t)\right]^2$$

where the regressor vector $\varphi(t)$ is given by

$$\varphi(t) = \bar{C}(q)\bar{e}(t) = \bar{C}(q)\left(T_d^{-1}(q) - 1\right)y(t) \tag{6}$$

In this case, the regressor matrix is defined as

$$\Phi(t) = \left[\varphi(1),\ldots,\varphi(N)\right]^{\mathrm{T}} \tag{7}$$

Therefore, the controller's parameters $\hat{\rho}$ can be estimated by minimizing the squares of the difference between $\hat{\rho}^{\mathrm{T}}\varphi(t)$ and $u(t)$ solving the following normal equation

$$\hat{\rho}(N) = \left[\sum_{k=1}^{N}\varphi(t)\varphi^{\mathrm{T}}(t)\right]^{-1}\left[\sum_{k=1}^{N}\varphi(t)u(t)\right] \tag{8}$$

where, N is the number of samples collected in the experiment, and $\hat{\rho}(N)$ is the parameters vector estimated up to the N-th sample. Now, it is possible to define the information matrix $P(N)$ as

$$P(N) = \left[\sum_{k=1}^{N}\varphi(t)\varphi^{\mathrm{T}}(t)\right] \tag{9}$$

When the assumption that the ideal controller belongs to the controller class is no longer guaranteed (also known as mismatched case) the regressor vector $\varphi(t)$ and the input signal $u(t)$ are filtered by a filter defined as $L(e^{j\omega}) = T_d(e^{j\omega})\left(1 - T_d(e^{j\omega})\right)$. This filter is applied to make the minimum of the cost functions close to each other. A more detailed explanation about the approaches of the VRFT to deal with the mismatched case, the noisy case, and other approaches can be found in [10].

III. Application of the criteria to the controller parameters estimation problem

A. Smallest singular value criterion

The application of this criterion to the controller parameters estimation problem can be justified by the fact that in the eigenparameter space the bias and variance of the estimated parameters $\hat{\rho}_V^{\mathrm{T}}(N)$ of the proposed problem behave as described in subsection II-A. To elucidate this fact 200 Monte Carlo experiments were performed in an open loop experiment with a step as system input. The parameters were estimated increasing the number of samples with the VRFT method. The bias and variance of the eigenparameters were calculated and are presented in Figure 1.

It is possible to see that the bias of the eigenparameter related to the smallest singular value $\hat{\rho}_{V2}(N)$ suffers the most with time, while the bias of $\hat{\rho}_{V1}(N)$ is approximately constant and close to zero. Whereas, the variance of the two parameters decreases with the number of samples.

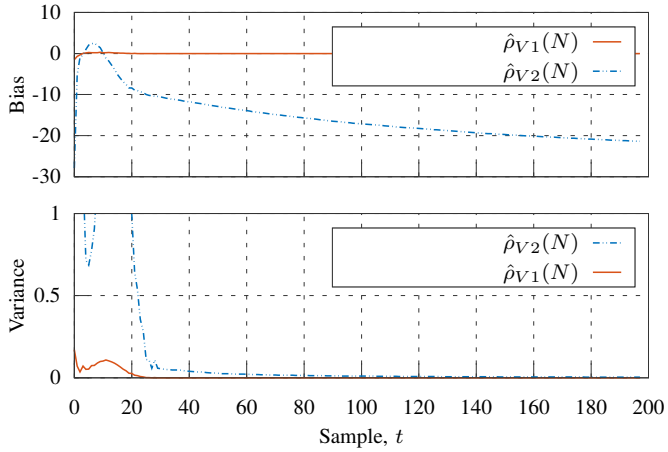

Fig. 1. Bias and variance of $\hat{\rho}_{V1}(N)$ solid line and $\hat{\rho}_{V2}(N)$ dashed line.

In order to apply the smallest singular value criterion to the parameters estimation problem and to allow a comparison between the methods, the following adjustment was made: the smallest singular value method was applied to the regressor matrix $\Phi(t)$ defined in (7) formed by the VRFT regressor vectors defined in (6) as

$$\underline{\sigma}^2(\Phi(t)) - \underline{\sigma}^2(\Phi(t-1)) < \eta_c$$

The informative subset is delimited while the inequality is not satisfied, this way, after the inequality is satisfied all remaining data are discarded.

B. Condition number criterion

In order to apply this criterion to the parameters estimation problem the condition number method was applied to the VRFT information matrix $P(N)$ in (9) as

$$\gamma(t) = \frac{\underline{\sigma}(P(N))}{\overline{\sigma}(P(N))} > \eta_n$$

The informative subset is defined while the inequality is satisfied, thus, the remaining data is discarded when the inequality is no longer satisfied.

C. Proposed criterion

The proposed criterion is still in the first steps and a proper mathematical formulation needs to be developed. This criterion was derived from graphical analysis of the reciprocal condition number $\gamma(t)$ and the estimated cost function $J_y(N)$. The $J_y(N)$ estimated is defined as

$$J_y(N) = \frac{1}{N_e}\sum_{t=1}^{N_e}\left(y(t,\hat{\rho}(N)) - y_d(t)\right)^2 \qquad (10)$$

where, N_e is the number of samples to calculate the cost function, $y_d(t)$ is the desired output, and $y(t,\hat{\rho}(N))$ is the obtained output with the controller gains $\hat{\rho}(N)$ estimated until the sample N, where N is increased one sample at a time.

The method searches for a good number of samples t_f to truncate the data

$$t_f = \arg\min_t \delta(t) \triangleq \gamma(t) - \hat{\gamma}(t) \qquad (11)$$

where, $\hat{\gamma}(t)$ is a polynomial approximation of $\gamma(t)$. Here, a sixth order polynomial was used to estimate $\hat{\gamma}(t)$. This choice was made because that was the order that best adjusted $\hat{\gamma}(t)$ to $\gamma(t)$.

In other words, the goal is to find the index of the sample where the data deviates the most from the fitted curve towards bellow. The sample where the minimum of $\delta(t)$ occurs corresponds to a point in the valley around the minimum of $J_y(N)$.

It is important to notice that one should ignore the first n_γ samples before looking for the minimum. Therefore, n_γ is a parameter that needs to be chosen by the designer.

IV. SIMULATION EXAMPLES

In this section the simulation experiments and the results obtained are presented. To each simulation example the criteria were applied as presented in section III. In all cases the output is affected by a colored noise $v(t)$ generated by a white Gaussian noise filtered by the following transfer function

$$H(q) = \frac{q^2 + 0.8q + 0.3}{q^2 - 0.8q}$$

The results obtained with each criterion are compared by the value obtained with the cost function $J_y(N)$ in (10). To compare numerically the obtained values it was calculated how many values for the $J_y(N)$ found remain in the smallest value of the cost function plus the difference between the smallest and the final value of the cost function divided by four, as

$$q = \min(J_y(N)) + \left(\frac{f(J_y(N)) - \min(J_y(N))}{4}\right) \qquad (12)$$

where $f(J_y(N))$ corresponds to the final value of the cost function.

A. Open loop case

In this simulation experiment 500 Monte Carlo simulations of an open loop experiment were performed. A step with 200 samples was used as input $u(t)$ for the system. The system transfer function $G(q)$ is given by

$$G(q) = \frac{0.42794(q - 0.9145)}{(q - 0.867)(q - 0.9587)} \qquad (13)$$

The ideal controller $C_i(q)$ is given by

$$C_i(q) = \frac{0.156q - 0.1454}{q - 1} \qquad (14)$$

As mentioned before, the parameters were estimated through the VRFT method increasing number of samples for each realization. The values used for the thresholds are: $\eta_c = 60\eta_{min}$ for the smallest singular value criterion, $\eta_n = 0.02$ for the reciprocal condition number criterion, and $\eta_\gamma = 15$ as initial start for the proposed criterion.

The confidence interval of the criteria is presented in Figure 2. As can be seen, regardless of the criterion used the parameters estimated are biased. This occurs because the simplest approach of the VRFT method is applied and the signals are affected by noise. Moreover, it is possible to observe that the bias of the parameters obtained applying the condition number criterion and the proposed criterion is almost equal, while the variance obtained with the proposed criterion is bigger then the variance obtained with the condition number criterion. In contrast, the smallest singular value criterion shows the highest values for bias and variance.

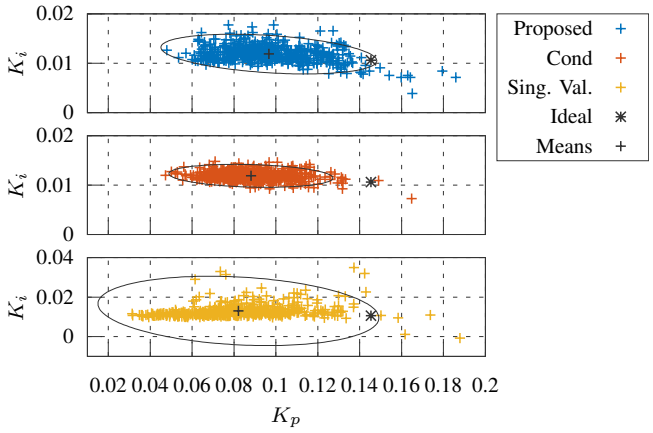

Fig. 2. Confidence interval of the criteria for the open loop experiment.

The numeric comparative was calculated as in (12) and the values found are presented in Table I. It is possible to observe that the proposed criterion and the reciprocal condition number criterion show better results than the smallest singular value criterion.

TABLE I
COMPARATIVE – EXPERIMENT 1

Criterion	Below q
Proposed	326 (65.2%)
Reciprocal condition number	294 (58.8%)
Smallest singular value	118 (23.6%)

The total mean square error (MSE) was also calculated, for each criterion, as

$$MSE = \sum_{i=1}^{n_p} \left(\text{bias}^2(\hat{\rho}_i) + \text{var}(\hat{\rho}_i) \right) \qquad (15)$$

where $\hat{\rho}$ is the estimated controller gains and n_p is the number of parameters. The obtained values are: 0.00289 for the proposed criterion, 0.00355 for the condition number criterion, and 0.00499 for the smallest singular value criterion.

The closed loop responses, obtained with one realization randomly chosen for each method, are presented in Figure 3. It is possible to see that the obtained responses are close to the desired one, different from the obtained response when the entire data set is used to estimate the parameters.

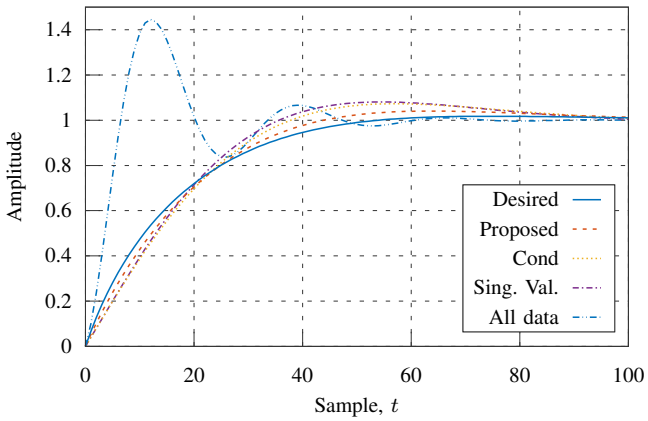

Fig. 3. Closed loop step responses for each criterion.

B. Closed loop case

In this case the same system transfer functions for $G(q)$ in (13) and $C_i(q)$ in (14) were used. The controller that was in the system when the experiments were performed is given by

$$C(q) = \frac{0.1469q - 0.1388}{q - 1}$$

In this case, a step sequence was used as input reference $r(t)$, and 500 Monte Carlo simulations were performed. Here, also for each realization, the parameters were estimated using the VRFT method with the number of samples increasing one sample each time. The values used for the thresholds are: $\eta_c = 80\eta_{min}$ for the smallest singular value criterion, $\eta_n = 0.02$ for the reciprocal condition number criterion, and $\eta_\gamma = 35$ as initial start for the proposed criterion. The confidence interval of the methods is presented in Figure 4. As can be seen, the bias obtained with the proposed criterion and the condition number criterion is almost equal. In this case, applying the smallest singular value we also obtained the highest values of the bias and variance for the estimated parameters.

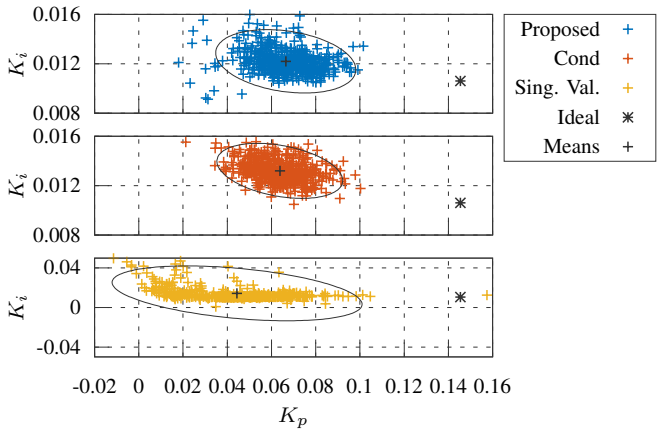

Fig. 4. Confidence interval of the criteria for the closed loop experiment.

The numeric comparative was calculated and is presented in Table II. In this case, the proposed criterion shows the best result.

TABLE II
COMPARATIVE – EXPERIMENT 2

Method	Below q
Proposed method	358 (71.6%)
Reciprocal condition number	139 (27.8%)
Smallest singular value	104 (20.8%)

The MSE was also calculated, for each criterion, as in (15) and the obtained values are: 0.00640 for the proposed criterion, 0.00681 for the condition number criterion, and 0.02295 for the smallest singular value criterion.

The cost function was calculated for four randomly selected realizations and is presented in Figure 5. Each cost function was calculated with increasing number of samples and the dots indicate the ending of the intervals returned by each criterion. It is clear that using only a subset of the data results in a smaller value for the cost function compared to the value found using the whole data set.

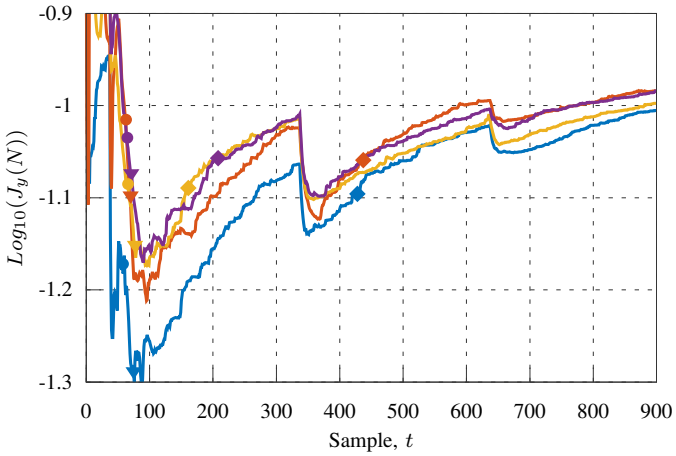

Fig. 5. Cost functions and points returned by each method of four randomly selected realizations. Each point corresponds to: ▼ the proposed criterion, ● the reciprocal condition number criterion, and ♦ the smallest singular value criterion.

C. Mismatched case

This simulation example comprises the case when the ideal controller does not belong to the controller class. In this case, also 500 Monte Carlo simulations were performed and a step was used as system input $u(t)$. The parameters were calculated using the VRFT approach for the mismatched case as described in subsection II-C. The system transfer function $G(q)$ is given by

$$G(q) = \frac{0.19963(q - 0.9783)}{(q - 0.8513)(q - 0.965)}$$

The controller class used to identify the controller's parameters is a PI, while the ideal controller is a PID given by

$$C_i(q) = \frac{0.1557q^2 - 0.1485q + 0.0354}{q^2 - q}$$

The values used for the thresholds are: $\eta_c = 1 \times 10^5 \eta_{min}$ for the smallest singular value criterion, $\eta_n = 0.002$ for the reciprocal condition number criterion, and $\eta_\gamma = 15$ as initial start for the proposed criterion. The numeric comparative was calculated and is presented in Table III. It is possible to see that for this case the obtained results are not so good. That can also be seen through the Bode diagrams in figures 6, 7 and 8.

TABLE III
COMPARATIVE – EXPERIMENT 3

Method	Below q
Proposed method	104 (20.8%)
Reciprocal condition number	85 (17%)
Smallest singular value	84 (16.8%)

In this case, the MSE was also calculated, for each criterion, as in (15) and the obtained values are: 0.00612 for the proposed criterion, 0.00646 for the condition number criterion, and 0.04655 for the smallest singular value criterion.

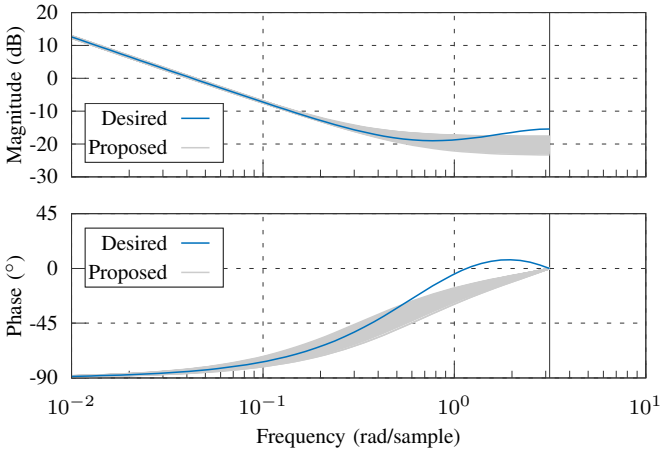

Fig. 6. Bode diagram of the controllers found in the Monte Carlo simulations using the proposed criterion.

V. CONCLUSIONS

In this work, some methods existent in the literature aiming at finding informative subsets from data gathered from normal operation routines, originally applied to the system identification framework, were applied to the controller's estimation problem. The methods applied are based on the smallest singular value criterion and the reciprocal condition number criterion. Moreover, a new criterion based on the reciprocal condition number was presented and applied to the same problem. Some simulation examples are also presented whose results indicate that using these criteria may improve the parameters estimation.

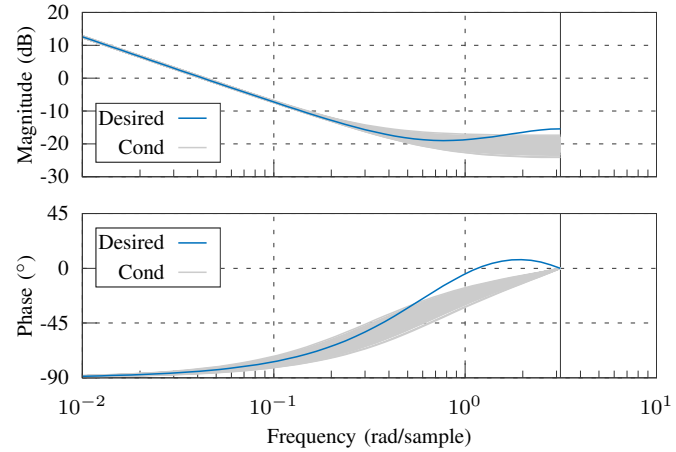

Fig. 7. Bode diagram of the controllers found in the Monte Carlo simulations using the condition number criterion.

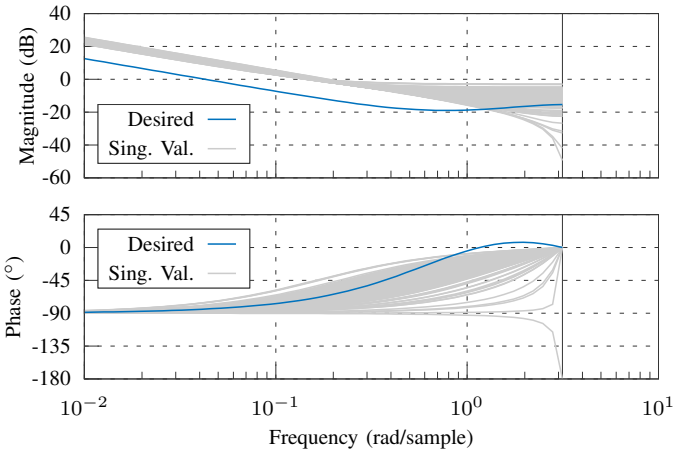

Fig. 8. Bode diagram of the controllers found in the Monte Carlo simulations using the smallest singular value criterion.

There are still some open issues as the development of a stronger mathematical proof of the applicability of the literature criteria to the controller's estimation problem. Also the development of a better mathematical formulation for the proposed criterion, and to extend these criteria to the multivariable case.

References

[1] P. Carrette, G. Bastin, Y. Genin, and M. Gevers, "Discarding data may help in system identification," *IEEE transactions on signal processing*, vol. 44, pp. 2300–2310, Sept. 1996.

[2] D. Peretzki, A. J. Isaksson, A. C. Bittencourt, and K. Forsman, "Data mining of historic data for process identification," in *Proc. of the AIChE Annual Meeting*, Minneapolis, USA, Oct. 2011, pp. 16–21.

[3] A. C. Bittencourt, A. J. Isaksson, D. Peretzki, and K. Forsman, "An algorithm for finding process identification intervals from normal operating data," *Processes*, vol. 3, pp. 357–383, May 2015.

[4] Y. A. W. Shardt and S. L. Shah, "Segmentation methods for model identification from historical process data," in *Proc. of the 19th IFAC World Congress*, Cape Town, South Africa, Aug 2014, pp. 2836–2841.

[5] Y. A. Shardt and B. Huang, "Data quality assessment of routine operating data for process identification," *Computers & Chemical Engineering*, vol. 55, pp. 19–27, apr 2013.

[6] D. Arengas and A. Kroll, "Searching for informative intervals in predominantly stationary data records to support system identification," in *Proc. of the XXVI International Conference on Information, Communication and Automation Technologies (ICAT)*, Sarajevo, Bosnia and Herzegovina, Oct 2017, pp. 1–6.

[7] ——, "A search method for selecting informative data in predominantly stationary historical records for multivariable system identification," in *Proc. of the 21st International Conference on System Theory, Control and Computing (ICSTCC)*, Sinaia, Romania, Nov 2017, pp. 100–105.

[8] J. Wang, J. Su, Y. Zhao, and D. Zhou, "Searching historical data segments for process identification in feedback control loops," *Computers & Chemical Engineering*, vol. 112, pp. 6–16, jan 2018.

[9] C. M. Holcomb and R. R. Bitmead, "Subspace identification with multiple data records: unlocking the archive," unpublished.

[10] A. S. Bazanella, L. Campestrini, and D. Eckhard, *Data-driven controller design: the H2 approach*. Netherlands: Springer Science & Business Media, 2011.