Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

Identification of pseudo-State Space Models for Batch Processes using Multivariate Statistical Methods ⋆

Eduardo B. López-Montero * **Ognjen Marjanovic** **

* *Control Systems Centre, School of Electrical and Electronic Engineering, The University of Manchester, Manchester, UK (e-mail: Eduardo.LopezMontero@postgrad.manchester.ac.uk)*
** *Control Systems Centre, School of Electrical and Electronic Engineering, The University of Manchester, Manchester, UK (e-mail: Ognjen.Marjanovic@manchester.ac.uk)*

Abstract: A new methodology to identify models in a pseudo-state space form for batch/fed-batch processes is proposed. The methodology employs historical data from previous batch runs, where a few intermittent measurements of product quality were made, and multivariate statistical methods in order to identify data-based models. Multivariate statistical methods, such as principal components analysis (PCA) and partial least squares (PLS), are being increasingly employed for batch processes model identification due to the advantages they offer over more difficult and time-consuming first-principle modelling techniques. In the proposed model identification approach, predictors are obtained employing PCA and PLS algorithms. Then, after a new vector of pseudo-states is defined, a pseudo-state space model is identified by performing an algebraic manipulation of the PCA and PLS statistical models. The ability of the pseudo-state space models to accurately predict future process variable trajectories is demonstrated by means of a simulation benchmark for penicillin production.

Keywords: Batch processes monitoring; Principal component analysis; Partial least squares; State space models.

1. INTRODUCTION

Continuous and discontinuous operations can be encountered in the chemical industry. Whilst continuous operations are mostly employed for high volume production, discontinuous ones (batch or semi-batch) are best suited for industry sectors focused on the manufacture of low-volume, high value added products such as specialty chemicals, pharmaceuticals, polymers, food, among others (Bonvin et al. (2006)). It is clear that batch processes constitute a very important part of the chemical industry. Furthermore, new market environment has generated an increase in the demand of low-volume high-added value products (Barbosa-Póvoa (2007)), which is the main reason why the development of batch process monitoring and control techniques has received considerable interest from both academia and industry.

The development of batch process models that can be used for monitoring and control represents an important challenge for process engineers, due to the complex characteristics of this type of discontinuous processes. Batch processes are distinguished by their finite duration, time-varying and non-linear dynamics, irreversible behaviour, and lack of equilibrium condition (Bonvin (1998)). Initial studies of batch processes were based on rigorous first-principles models. However, the identification of such type of models is very time consuming. Furthermore, when this type of models is being developed it is common practice to make simplifying assumptions, due to the difficulty that the characterization of a complex chemical phenomena represents (Luyben (2007)). In practice, such assumptions could make the model unreliable when used for predictions purposes.

In recent years, multivariate statistical methods have been increasingly used to identify data-based models in order to monitor and control batch or semi-batch processes; the latter can be attributed to the advantages these methods offer over the ones requiring deep theoretical understanding of the process. Data-based modelling does not require detailed a-priori knowledge of the process, models are relatively easy to identify and keep up to date. Amongst the data-driven methodologies used to identify batch process models for monitoring and control purposes, multi-way principal component analysis (MPCA) and multi-way partial least squares (MPLS), which were proposed by MacGregor and co-workers (Kresta et al. (1991); Nomikos and MacGregor (1994)), have received particular attention from researchers. Recently, these techniques have been further investigated resulting in the development of methodologies able to identify models for control purposes. In Golshan et al. (2010) a multivariate model predictive control (MPC) is proposed based on a multi-phased principal component analysis (PCA) model. This methodology has been further studied in Golshan et al. (2011), where

⋆ This research project is funded by the Mexican Science and Technology Council (CONACyT).

different modelling alternatives were investigated. First a model for the batch process is identified using historical data, then a trajectory tracking controller is designed around such PCA model. The controller has most of the characteristics of a standard MPC. However, due to the nature of the model, constraints to the manipulated variables can not be included straightforwardly; furthermore, during a new batch run, measurements of process variables must be appropriately scaled before using them for prediction purposes. These are reasons why the MPC controller has to be specifically designed for this type of data-based models, therefore standard MPC cannot be formulated using such models. Furthermore, the technique proposed in Golshan et al. (2010) is only capable of estimating future values of the readily measured variables, it does not provide predictions of batch product quality. In Wan et al. (2012) a robust control methodology is proposed, where a partial least squares (PLS) model for batch end-product quality is identified and then a tailored MPC controller is used to achieve batch end-quality specifications. However, the identified PLS model is only able to predict product quality at the batch end-point rather than during the batch progression. Therefore, a controller is designed specifically for this type of model. Such controller has MPC characteristics, but due to the way the PLS model is constructed standard MPC cannot be formulated.

This paper presents an alternative modelling technique based on PCA and PLS algorithms in order to identify models for batch or semi-batch processes. The proposed model identification methodology assumes that a few intermittent measurements of batch product quality were taken during the batch runs that form the training data. Then the data is arranged using the unfolding approach proposed in Marjanovic et al. (2006). Therefore, batch quality predictions can be made throughout the duration of the batch. The models obtained using the methodology proposed in this paper are in a general state space form, and apart from being able to deliver accurate predictions of process variable values (including quality related ones) during the batch operating time, they offer the possibility to be integrated into standard MPCs for trajectory tracking of batch product quality. The remainder of the paper is structured as follows. In Section 2 the PCA and PLS algorithms are briefly described. Then, the pseudo-state space model identification methodology based on multivariate statistical methods and intermittent measurements is explained in Section 3. Some results regarding the prediction capabilities of the proposed modelling methodology are shown in Section 4. And finally, some concluding remarks and intended future work are presented in Section 5.

2. MULTIVARIATE STATISTICAL METHODS

Multivariate statistical methods such as PCA and PLS have been successfully used in continuous processing applications, where data matrices are two-dimensional arrays. However, the historical data-set collected from a batch process forms a three-dimension matrix X consisting of I batches, J process variables [1], and K sample instants, as shown in Fig 1. In order to handle three-dimensional arrays, MPCA and MPLS are used, which are capable of handling data matrices with three dimensions after the data has undergone through an unfolding step; in fact, MPCA and MPLS are equivalent to carrying out standard PCA and PLS on larger two-dimensional arrays. There are several ways to unfold the data collected from a batch process, with batch wise unfolding (BWU) considered the most logical approach for modelling differences among the batches (Golshan et al. (2011)). The BWU approach is depicted in Fig. 1 where the original matrix $X\,(I \times J \times K)$ is unfolded into a matrix with two dimensions $X\,(I \times JK)$, in which each row corresponds to the information from each individual batch run, and all the process variables at different sample instants are put beside each other.

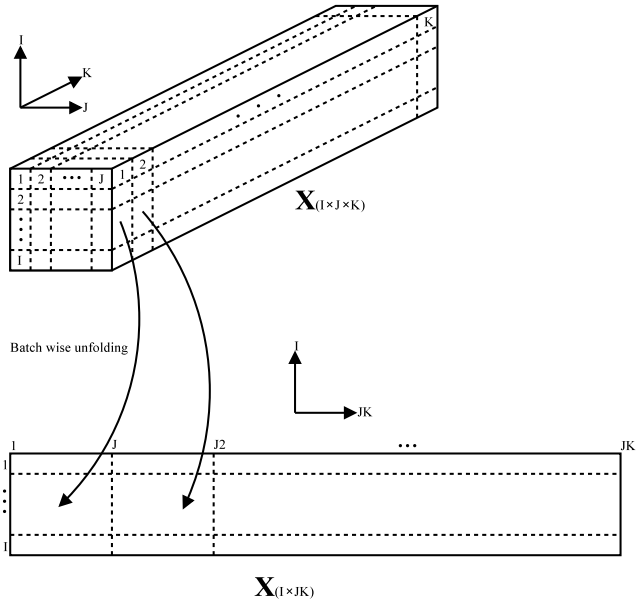

Fig. 1. Batch wise unfolding of original data-set.

2.1 Multi-way principal component analysis

MPCA is a method consisting of applying PCA to the unfolded batch process data [2]. Such method is used for projecting the original data onto a score space T with reduced dimensions:

$$X = TP^T + E \tag{1}$$

where $T \in \mathbb{R}^{I\times n_{lv}}$, $P \in \mathbb{R}^{JK\times n_{lv}}$, and $E \in \mathbb{R}^{I\times JK}$ are the scores, loadings and residuals matrices respectively, n_{lv} represents the number of latent variables retained that account for most of the data variability, this value can be obtained through cross-validation (Tan et al. (2012)). If an appropriate n_{lv} is selected, the residual matrix E contains negligible information and can be discarded.

2.2 Multi-way partial least squares

In order to identify a MPLS model, batch process data is divided into cause/input variables and effect/output variables. The input matrix usually corresponds to the unfolded process variables $X\,(I \times JK)$, and the output

[1] Total number of process variables $J = n_x + n_u$, where n_x is the number of readily measured variables, and n_u is the number of manipulated variables.

[2] Prior to performing PCA or PLS, the data is scaled to zero mean and unit variance.

matrix $Y\,(I \times n_y)$ corresponds to measurements of batch product quality, where n_y is the number of quality related outputs. Note that such measurements are usually taken at the end of each batch run. Performing PLS on these matrices also results in a reduced dimension model of the form:

$$X = TP^T + E \tag{2}$$
$$Y = UQ^T + F \tag{3}$$

where $T \in \mathbb{R}^{I \times n_{lv}}$, $P \in \mathbb{R}^{JK \times n_{lv}}$, and $E \in \mathbb{R}^{I \times JK}$ are the input scores, loadings and residuals matrices, respectively; similarly, $U \in \mathbb{R}^{I \times n_{lv}}$, $Q \in \mathbb{R}^{n_y \times n_{lv}}$, and $F \in \mathbb{R}^{I \times n_y}$ are the output scores, loadings and residuals matrices, respectively. The input and output scores are related by a diagonal matrix $\beta \in \mathbb{R}^{n_{lv} \times n_{lv}}$ so that $U = T\beta$. The non-linear iterative partial least squares (NIPALS) regression algorithm is normally used to obtain the PLS model (Wold et al. (1987)), in this algorithm an additional weighting matrix $W \in \mathbb{R}^{JK \times n_{lv}}$ is used to calculate the scores $T = XW\left(P^TW\right)^{-1}$. In practice, the PLS model is often expressed as a predictive model directly relating the input and output variables:

$$Y = X \underbrace{W\left(P^TW\right)^{-1}\beta Q^T}_{\Theta} + F^* \tag{4}$$

where F^* is the residuals matrix containing negligible information if an appropriate n_{lv} were selected.

3. PSEUDO-STATE SPACE MODEL IDENTIFICATION

The proposed methodology for identifying models for batch processes in a general state space form is based on the data arrangement proposed by Marjanovic et al. (2006), and on PCA/PLS algorithms. After the data-based models have been obtained and a vector of pseudo-states has been defined, algebraic manipulation is carried out on the PCA and PLS models in order to obtain a model in a standard state space representation.

3.1 Data arrangement

The method for identifying statistical models using intermittent measurements consists in a re-arrangement of the unfolded data. Pseudo-batches are created at those intermittent measurement instants and they are aligned toward their end-points. Then, a modelling window size (K_w) is selected and all the information outside such window is discarded. The windows size is equal to the size of the smallest pseudo-batch.

In order to illustrate the new data alignment Fig. 2 is presented, where data from two different batch runs is depicted; during each batch run three measurements of product quality were taken, therefore a total of six pseudo-batches ($I_w = 6$) are created and they are aligned toward their end-points as shown in Fig. 2. Afterwards, K_w is selected to be equal to the number of sample instants contained in the smallest pseudo-batch. As a result, two new input and output matrices are formed: $X_w\,(I_w \times JK_w)$ and $Y_w\,(I_w \times n_y)$. For a more detailed description of the data arrangement the reader is referred to Marjanovic et al. (2006).

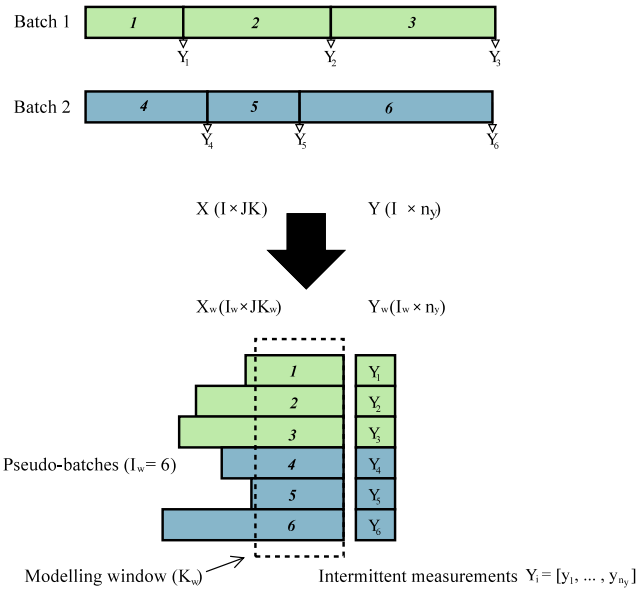

Fig. 2. Batch wise unfolding of original data containing intermittent measurements.

In order to identify the pseudo-state space model, an additional input matrix $\left(X_{ss} \in \mathbb{R}^{I_w \times J(K_w - 1)}\right)$ is formed. This matrix is obtained by removing the first J columns of X_w. The new input and output matrices have the form:

$$X_w = \begin{bmatrix} x_{1,1}^T & u_{1,1}^T & \cdots & x_{1,K_w}^T & u_{1,K_w}^T \\ \vdots & & & & \vdots \\ x_{I_w,1}^T & u_{I_w,1}^T & \cdots & x_{I_w,K_w}^T & u_{I_w,K_w}^T \end{bmatrix} \tag{5}$$

$$X_{ss} = \begin{bmatrix} x_{1,2}^T & u_{1,2}^T & \cdots & x_{1,K_w}^T & u_{1,K_w}^T \\ \vdots & & & & \vdots \\ x_{I_w,2}^T & u_{I_w,2}^T & \cdots & x_{I_w,K_w}^T & u_{I_w,K_w}^T \end{bmatrix} \tag{6}$$

$$Y_w = \begin{bmatrix} y_{1,1} & \cdots & y_{1,n_y} \\ \vdots & & \vdots \\ y_{I_w,1} & \cdots & y_{I_w,n_y} \end{bmatrix} \tag{7}$$

where $x \in \mathbb{R}^{n_x \times 1}$ is the column vector of the readily measured variables, and $u \in \mathbb{R}^{n_u \times 1}$ is the column vector of the manipulated variables.

3.2 Model identification

Before performing PCA or PLS analysis, each column of matrices X_w, X_{ss} and Y_w is scaled to zero mean and unit variance. The mean (μ) and standard deviation (σ) row vectors used for scaling the data arrays are:

$$\mu_{X_w} = \left[\, \bar{m}_1 \ \cdots \ \bar{m}_{JK_w} \,\right] \in \mathbb{R}^{1 \times JK_w} \tag{8}$$
$$\sigma_{X_w} = \left[\, std_1 \ \cdots \ std_{JK_w} \,\right] \in \mathbb{R}^{1 \times JK_w} \tag{9}$$
$$\mu_{Y_w} = \left[\, \bar{m}_1 \ \cdots \ \bar{m}_{n_y} \,\right] \in \mathbb{R}^{1 \times n_y} \tag{10}$$
$$\sigma_{Y_w} = \left[\, std_1 \ \cdots \ std_{n_y} \,\right] \in \mathbb{R}^{1 \times n_y} \tag{11}$$

where $\bar{m}_n$ and std_n are the mean and standard deviation, respectively, of the n-th column of the corresponding matrix. PCA is then applied to matrix X_w, and PLS is applied to matrices X_{ss} and Y_w in order to identify the statistical models:

$$X_w = TP^T + E \tag{12}$$

$$Y_w = X_{ss}\Theta + F \tag{13}$$

After the PCA and PLS models have been obtained, a pseudo-states vector ($\mathbf{z}$) is to be defined. Such vector at sample instant k has the form:

$$\mathbf{z}_k = \left[\, x_{k-K_{ss}+1}^T \;\; u_{k-K_{ss}+1}^T \;\cdots\; x_k^T \;\; u_k^T \,\right]^T \tag{14}$$

where $K_{ss} = K_w - 1$. Then, in order to estimate the future values of the readily measured variables, x, at sample instant $k+1$, a data imputation algorithm called projection to the model plane (PMP), presented in Nelson et al. (1996), is used along with the PCA model depicted in equation (12). First, the loadings matrix P is grouped into two parts, corresponding to the known, P^*, and unknown, $P^\sharp$, information:

$$P^* = \begin{bmatrix} p_1 \\ \vdots \\ p_{JK_{ss}} \\ p_{JK_{ss}+n_x+1} \\ \vdots \\ p_{JK_w} \end{bmatrix} \in \mathbb{R}^{(JK_{ss}+n_u)\times n_{lv}} \tag{15}$$

$$P^\sharp = \begin{bmatrix} p_{JK_{ss}+1} \\ \vdots \\ p_{JK_{ss}+n_x} \end{bmatrix} \in \mathbb{R}^{n_x \times n_{lv}} \tag{16}$$

where p_m represents the m-th row vector of the loadings matrix P. Using PMP, future values of the readily measured variables can be calculated as follows:

$$x_{k+1} = \underbrace{P^\sharp \left(P^{*T}P^*\right)^{-1} P^{*T}}_{\Gamma} \begin{bmatrix} \mathbf{z}_k \\ u_{k+1} \end{bmatrix} \tag{17}$$

where $\Gamma \in \mathbb{R}^{n_x \times (JK_{ss}+n_u)}$. In equation (17) Γ can be divided in two matrices so that $x_{k+1} = [\Gamma_z \; \Gamma_u] \begin{bmatrix} \mathbf{z}_k \\ u_{k+1} \end{bmatrix}$; where $\Gamma_z \in \mathbb{R}^{n_x \times JK_{ss}}$ and $\Gamma_u \in \mathbb{R}^{n_x \times n_u}$. Next, matrices $\overline{A}$ and $\overline{B}$ can be formed:

$$\overline{A} = \begin{bmatrix} \mathbf{0}\left(J\left(K_{ss}-1\right)\times J\right) & \mathbf{I}_{J(K_{ss}-1)} \\ \Gamma_z & \\ \mathbf{0}\left(n_u \times JK_{ss}\right) & \end{bmatrix} \tag{18}$$

$$\overline{B} = \begin{bmatrix} \mathbf{0}\left(J\left(K_{ss}-1\right)\times n_u\right) \\ \Gamma_u \\ \mathbf{I}_{n_u} \end{bmatrix} \tag{19}$$

In equation (18) and equation (19), $\mathbf{0}\left(m \times n\right)$ represents a matrix of zeros with m rows and n columns. Similarly, $\mathbf{I}_i$ represents an identity matrix of size i. Utilising matrices defined in equations (13), (18) and (19), the state space model used to compute the future pseudo-state ($\mathbf{z}_{k+1} \in \mathbb{R}^{JK_{ss}\times 1}$) and output ($\mathbf{y}_k \in \mathbb{R}^{n_y \times 1}$) column vectors can be defined as follows:

$$\begin{aligned}\mathbf{z}_{k+1} = \big\{&\overline{A}\left[\left(\mathbf{z}_k - \mu_1^T\right) \oslash \sigma_1\right] + \\ &\overline{B}\left[\left(u_{k+1} - \mu_2^T\right) \oslash \sigma_2\right]\big\} \otimes \sigma_3^T + \mu_3^T\end{aligned} \tag{20}$$

$$\mathbf{y}_k = \left\{\Theta^T\left[\left(\mathbf{z}_k - \mu_3^T\right) \oslash \sigma_3\right]\right\} \otimes \sigma_{Y_w}^T + \mu_{Y_w}^T \tag{21}$$

where $\oslash$ and $\otimes$ represents the Hadamard division and multiplication, respectively. In equation (20) and equation (21), standard deviation ($\sigma_{1:3}$) and mean ($\mu_{1:3}$) row vectors are constructed as follows:

$$\sigma_1 = \sigma_{X_w}\left(1 : JK_{ss}\right) \tag{22}$$

$$\sigma_2 = \sigma_{X_w}\left(JK_w - n_u + 1 : JK_w\right) \tag{23}$$

$$\sigma_3 = \sigma_{X_w}\left(J+1 : JK_w\right) \tag{24}$$

$$\mu_1 = \mu_{X_w}\left(1 : JK_{ss}\right) \tag{25}$$

$$\mu_2 = \mu_{X_w}\left(JK_w - n_u + 1 : JK_w\right) \tag{26}$$

$$\mu_3 = \mu_{X_w}\left(J+1 : JK_w\right) \tag{27}$$

Note that the scaling vectors (σ, μ) can be incorporated into the state space model in equation (20) and equation (21). Resultant state-space model matrices A, B, C, μ_z, and μ_y are then given as follows:

$$A = \overline{A} \otimes \left[\sigma_3^T \cdot \left(\mathbf{1} \oslash \sigma_1\right)\right] \tag{28}$$

$$B = \overline{B} \otimes \left[\sigma_3^T \cdot \left(\mathbf{1} \oslash \sigma_2\right)\right] \tag{29}$$

$$C = \Theta^T \otimes \left[\sigma_{Y_w}^T \cdot \left(\mathbf{1} \oslash \sigma_3\right)\right] \tag{30}$$

$$\mu_z = \mu_3^T - A\mu_1^T - B\mu_2^T \tag{31}$$

$$\mu_y = \mu_{Y_w}^T - C\mu_3^T \tag{32}$$

where $\mathbf{1}$ represents a row vector of ones; for equation (28) and equation (30) the size of such vector is $(1 \times JK_{ss})$, and for equation (29) the size of $\mathbf{1}$ is $(1 \times n_u)$. Therefore, the pseudo-state space model based on multivariate statistical methods is:

$$\mathbf{z}_{k+1} = A\mathbf{z}_k + Bu_{k+1} + \mu_z \tag{33}$$

$$\mathbf{y}_k = C\mathbf{z}_k + \mu_y \tag{34}$$

The structure of the model depicted in equations (33) and (34) can be related to subspace model identification (SMI) applied to continuous processes. Therefore, it should be possible to utilize the modelling procedure presented in this paper to identify state-space models of continuous processes, which would exploit the ability of PLS technique to handle highly correlated data sets. A work concerning the application of PLS-based models within the MPC control framework to continuous processes is presented in Laurí et al. (2013).

4. CASE STUDY

In this section, a benchmark simulation of a fed-batch process is used to demonstrate the ability of the proposed modelling approach to identify models that can make accurate predictions of batch product quality. The model validation was based on statistical indices for comparing predicted variable values against the actual ones.

4.1 Process description and model identification

The case study used was developed by Birol et al. (2002), which corresponds to a benchmark simulation for fed-batch fermentation of penicillin and is based upon a series of detailed mechanistic models that describe the fermentation process. Although the actual penicillin simulator

considers multitude of various process variables, many of these cannot be continuously measured in most real-world applications. Therefore, in this case study only the following variables are assumed to be measured hourly in order to ensure that the case study is realistic:

(1) Aeration rate.
(2) Agitation power.
(3) Substrate feed temperature.
(4) Carbon dioxide (CO_2) concentration.
(5) Dissolved oxygen (DO) concentration.
(6) Culture volume.
(7) pH.
(8) Fermenter temperature.

These eight process variables form the vector x, and the only manipulated variable u is the substrate feed rate. The quality related output variable was considered to be the biomass concentration, which can only be measured intermittently throughout the batch operating time.

Data from 30 batches was collected for model building, with each batch having a duration of $K = 200$ hours, the process variables were measured hourly and filtered pseudo random binary signals (PRBS) were superimposed on the nominal substrate feed rate of 0.045 l/hr, as well as on the aeration rate and agitation power in order to excite process dynamics. It was further assumed that a few biomass concentration measurements were taken during each batch run. A single sample was randomly taken between the 45^{th} and 55^{th} hour, another one between the 95^{th} and 105^{th} hour, a third one between the 145^{th} a 155^{th} hour; and a last one by the end of each batch cycle. Therefore, four measurements were taken during each batch run, resulting in the total of 120 pseudo batches, which were created and aligned according to the procedure depicted in Fig. 2. The length for the modelling window was $K_w = 45$ corresponding to the size of the smallest pseudo-batch. Afterwards, a pseudo-state space model is identified based on PCA and PLS, following the methodology described in Section 3.

4.2 Model validation

The performance indices used to analyse the model accuracy were the mean of absolute percentage error (MAPE) and the R^2 statistic:

$$\text{MAPE} = \frac{100}{K - K_w + 1} \sum_{k=K_w}^{K} \left| \frac{\widehat{Y}_k - Y_k}{Y_k} \right| \tag{35}$$

$$R^2 = 1 - \frac{\sum_{k=K_w}^{K} \left(\widehat{Y}_k - Y_k\right)^2}{\sum_{k=K_w}^{K} \left(\widehat{Y}_k - \overline{Y}\right)^2}, \textit{where } \overline{Y} = \frac{\sum_{k=K_w}^{K} Y_k}{K - K_w + 1} \tag{36}$$

MAPE provides the absolute error in terms of percentage, where error is defined as the difference between the predicted ($\widehat{Y}$) and the actual (Y) values. The R^2 statistic is used to asses the amount of variability accounted for by the model and it ranges between 0 and 1, with the value of 1 indicating a perfect fit.

In order to demonstrate the prediction capabilities of the identified model, 50 new batches were simulated and the differences between the estimated and actual values were analysed. During the simulations a white noise with a signal to noise ratio [3] (SNR) equal to 60 dB was added to the measurements. For each batch run the first and only decision point was at $k = 45$, from this instant (assuming that the future values of the manipulated variable were known) the future process variable trajectories were predicted using the identified state space model. Fig. 3 shows the MAPE obtained from biomass predictions in histogram form. By observing the histogram it can be noted that the percentage of error is approximately between 0.2% and 1.4% for all the simulated batches, and that for the majority of them the error was actually less than 1%. These results indicate that the biomass predictions obtained with the model were very close to the actuals values.

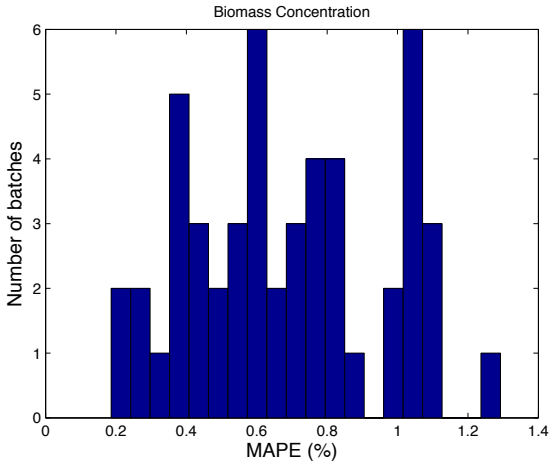

Fig. 3. Histogram of the MAPE from 50 batches.

Fig. 4 presents the histogram of the R^2 statistic obtained from predictions of the biomass concentration, where the values are expressed in terms of percentage. By inspecting this figure it can be seen that the percentage of variability accounted for was at least 95.5% approximately. Furthermore, by paying closer attention to Fig. 4 it can be seen that most of the variability accounted for by the model during the 50 batches was actually greater than 97%, only eight batches had R^2 values between 95.5% and 97%. Therefore, from the information presented in Fig. 3 and Fig. 4 it can be concluded that the model is able to accurately predict the biomass concentration.

Additionally, the plot shown in Fig. 5 is included to show the amount of percentage error obtained for each readily measured process variable during the 50 batch simulations. In Fig. 5 each box corresponds to a readily measured variable according to the list presented at the beginning of this section, e.g. variable 4 is CO_2 concentration. For each box, the central red mark represents the median, the edges of the boxes are the 25^{th} and 75^{th} percentiles and the whiskers extend to the most extreme data points. By inspecting this plot it can be noted that the percentage of error was less than 3.5% for all the variables. Furthermore, from a closer look to Fig. 5 it can be seen that only the agitation power (variable 2) had estimation errors between

[3] $\text{SNR} = 20 \log \left(\frac{\text{RMS}_{signal}}{\text{RMS}_{noise}} \right) dB$

0.5% and 3.5% approximately. For the rest of the variables the error was below 1%.

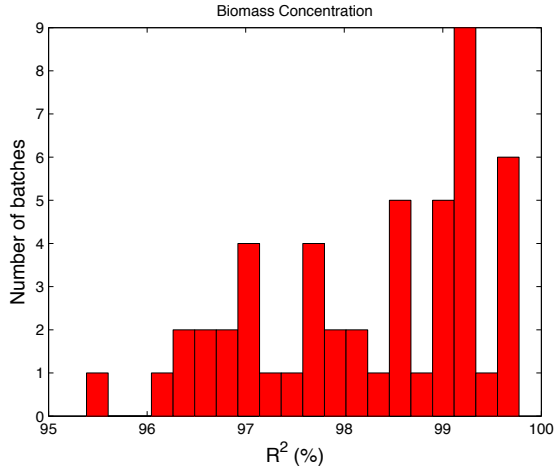

Fig. 4. Histogram of R^2 from 50 batches.

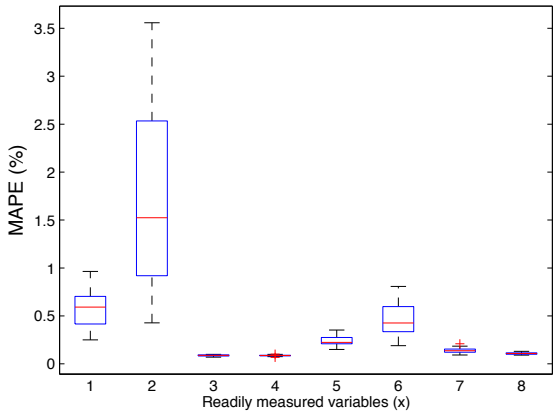

Fig. 5. Average MAPE for each readily measured variable.

5. CONCLUSIONS

In this paper a new methodology to identify batch processes models in a general state space form was proposed, which represents a practical way to obtain prediction models using multivariate statistical methods. From the case study employed it was demonstrated that the model obtained using the proposed methodology was able to give accurate predictions throughout the batch operating time. Predictions were obtained for both the continuously measured process variables and, more importantly, the quality related variables. This could facilitate design of controllers for trajectory tracking of batch product quality. Since the models obtained with the methodology presented in this paper are in a general state space form, standard MPC formulations could be used to design batch process controllers, opposed to the common practice of designing a predictive controller specifically for the PCA or PLS models when such statistical techniques are used for model identification. Hence, future work will be carried out where the proposed modelling approach will be used within general MPC architecture for developing trajectory tracking controllers for batch/fed-batch product quality.

ACKNOWLEDGEMENTS

The authors would like to acknowledge The Process Modelling, Monitoring, and Control Research Group at Illinois Institute of Technology (IIT) who generously provided the source code for their Pensim simulator.

REFERENCES

Barbosa-Póvoa, A.P. (2007). A critical review on the design and retrofit of batch plants. *Computers & Chemical Engineering*, 31(7), 833–855.

Birol, G., Ündey, C., and Çinar, A. (2002). A modular simulation package for fed-batch fermentation: penicillin production. *Computers & Chemical Engineering*, 26, 1553–1565.

Bonvin, D. (1998). Optimal operation of batch reactors - a personal view. *Journal of Process Control*, 8(5-6), 355–368.

Bonvin, D., Srinivasan, B., and Hunkeler, D. (2006). Control and Optimization of Batch Processes - Improvement of process operation in the production of specialty chemicals. *IEEE Control Systems Magazine*, 34–45.

Golshan, M., MacGregor, J.F., Bruwer, M.J., and Mhaskar, P. (2010). Latent Variable Model Predictive Control (LV-MPC) for trajectory tracking in batch processes. *Journal of Process Control*, 20(4), 538–550.

Golshan, M., MacGregor, J.F., and Mhaskar, P. (2011). Latent variable model predictive control for trajectory tracking in batch processes: Alternative modeling approaches. *Journal of Process Control*, 21(9), 1345–1358.

Kresta, J.V., Macgregor, J.F., and Marlin, T.E. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, 69, 35–47.

Laurí, D., Sanchis, J., Martínez, M., and Hilario, a. (2013). Latent variable based model predictive control: Ensuring validity of predictions. *Journal of Process Control*, 23(1), 12–22.

Luyben, W.L. (2007). *Chemical Reactor Design and Control*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Marjanovic, O., Lennox, B., Sandoz, D., Smith, K., and Crofts, M. (2006). Real-time monitoring of an industrial batch process. *Computers & Chemical Engineering*, 30(10-12), 1476–1481.

Nelson, P.R.C., Taylor, P.A., and MacGregor, J.F. (1996). Missing data methods in PCA and PLS : Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, 35, 45–65.

Nomikos, P. and MacGregor, J.F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8), 1361–1375.

Tan, C., Wu, T., Xu, Z., Li, W., and Zhang, K. (2012). A simple ensemble strategy of uninformative variable elimination and partial least-squares for near-infrared spectroscopic calibration of pharmaceutical products. *Vibrational Spectroscopy*, 58, 44–49.

Wan, J., Marjanovic, O., and Lennox, B. (2012). Disturbance rejection for the control of batch end-product quality using latent variable models. *Journal of Process Control*, 22(3), 643–652.

Wold, S., Geladi, P., Esbensen, K., and Öhman, J. (1987). Multi-way principal components and PLS-analysis. *Journal of Chemometrics*, 1, 41–56.