# Multivariate Analysis and Monitoring of Sequencing Batch Reactor Using Multiway Independent Component Analysis

ChangKyoo Yoo[*] and Peter A. Vanrolleghem
BIOMATH, Ghent University, Coupure Links 653, B-9000 Gent, Belgium

## Abstract

This contribution describes the monitoring on a pilot-scale sequencing batch reactor (SBR) using a batchwise multiway independent component analysis method (MICA) which can extract meaningful hidden information from non-Gaussian data. Given that independent component analysis (ICA) is superior to principal component analysis (PCA) to extract features from non-Gaussian data sets, the use of ICA may improve monitoring performance. The monitoring results of a pilot-scale SBR for biological wastewater treatment showed the power and advantages of MICA monitoring in comparison to conventional monitoring methods.

**Keywords:** Batch monitoring, Multivariate statistical process monitoring (MSPM), Multiway independent component analysis (MICA), Sequencing batch reactor (SBR)

## 1. Introduction

Sequencing batch reactor (SBR) processes have demonstrated their efficiency and flexibility in the treatment of wastewaters with high concentrations of nutrient, nitrogen, phosphorous, and toxic compounds from domestic and industrial sources. A SBR has a unique cyclic batch operation, usually with five well-defined phases: fill, react, settle, draw and idle. Most of the advantages of SBR processes may be attributed to their single-tank designs and the flexibility that allows them to meet many different treatment objectives, and which is derived from the possibility of adjusting the duration of the different phases. But the SBR process is highly nonlinear, time-varying and subject to significant disturbances like hydraulic changes, composition variations and equipment failures. Small changes in concentrations or flows can affect effluent quality and microorganism growth. However, treatment performance, the key indicator of process performance, is often only examined off-line in a laboratory. Even though operators are aware that there are some problems in treatment performance, they cannot quickly find out or predict what the causes are and when the problems will occur because most batch processes are run without any effective form of real-time on-line monitoring. Therefore, multivariate analysis and process monitoring of SBR are crucial to detect faults that can be corrected prior to completion of the batch or can be corrected in subsequent batches

[*] Author to whom correspondence should be addressed : ChangKyoo.Yoo@biomath.UGent.be

because it may take several days, week or ever months for the biological process to recover from abnormal operation (Lee and Vanrolleghem, 2003).

Multiway principal component analysis (MPCA) developed by Nomikos and MacGregor (1994) has been shown to be a powerful monitoring tool in many industrial batch processes. However, it has the shortcoming that the measurement variables of the batch process should be normally distributed. In this work, it is shown that multiway independent component analysis suggested by Yoo et al. (2003) can be used to overcome this drawback and obtain better monitoring performance.

## 2. Theory

### 2.1 Independent component analysis (ICA)

What distinguishes ICA from other methods is that it looks for components that are both statistically independent and non-Gaussian. PCA is a dimensionality reduction technique in terms of capturing the variance of the data which is capable of extracting uncorrelated latent variables from correlated data, while ICA is designed to separate the independent components (ICs) that are independent and constitute the observed variables. Furthermore, PCA can only impose independence up to second order statistics information (mean and variance) while constraining the direction vectors to be orthogonal, whereas ICA has no orthogonality constraint and also involves higher-order statistics (Hyvärinen et al., 2001). Hence, ICA may reveal more useful information in the non-Gaussian data than PCA (Hyvärinen et al., 2001).

In the ICA algorithm, it is assumed that $d$ measured variables $x_1, x_2, \ldots, x_d$ can be expressed as linear combinations of $m(\leq d)$ unknown independent components $s_1, s_2, \ldots, s_m$. The relationship between them is given by

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)] \in R^{d \times n}$ is the data matrix (in contrast to PCA, ICA employs the transposed data matrix.), $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_\mathrm{m}] \in R^{d\times m}$ is the unknown mixing matrix, $\mathbf{S} = [\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(n)] \in R^{m\times n}$ is the independent component matrix, $\mathbf{E} \in R^{d\times n}$ is the residual matrix, and $n$ is the number of samples. Here, we assume $d \geq m$ (when $d=m$, the residual matrix, $\mathbf{E}$, becomes the zero matrix). The basic problem of ICA is to estimate both the mixing matrix $\mathbf{A}$ and the independent components $\mathbf{S}$ from only the observed data $\mathbf{X}$. Alternatively, one could define the objective of ICA as follows: to find a demixing matrix $\mathbf{W}$ whose form is such that the rows of the reconstructed matrix $\hat{\mathbf{S}}$, given as

$$\hat{\mathbf{S}} = \mathbf{WX} \tag{2}$$

become as independent of each other as possible (Hyvärinen et al., 2001).

### 2.2 Multiway Independent Component Analysis (MICA)

The monitoring method based on MICA is similar to that based on MPCA. MICA is equivalent to performing ICA on a large two-dimensional matrix $\mathbf{X}$ constructed by batchwise unfolding the three-way data matrix $\underline{\mathbf{X}}$. MICA decomposes the three-way array $\underline{\mathbf{X}}$ into a summation of the product of independent vectors $\mathbf{s}_r$ and loading matrices

$\mathbf{A_r}$ plus a residual array $\underline{\mathbf{E}}$ so that the ICs $\mathbf{s}$ become as independent of each other as possible:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{s}_r \otimes \mathbf{A}_r + \underline{\mathbf{E}} = \sum_{r=1}^{R} s_r \mathbf{a}_r^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \tag{3}$$

where ⊗ denotes the Kronecker product ($\underline{\mathbf{X}} = \mathbf{s} \otimes \mathbf{A}$ is $\underline{X}(i,j,k) = s(i)A(j,k)$) and $R$ denotes the number of ICs retained. The $\mathbf{S}$ and $\mathbf{A}$ matrices in Eq. (3) can be equivalent to the loading matrix and score matrices by analogy with MPCA, *i.e.* $\mathbf{S}$ can be regarded as the score matrix $\mathbf{T}$, and $\mathbf{A}$ can be treated as the loading matrix $\mathbf{P}$. The *i*th elements of the independent vector $\mathbf{s}$ correspond to the *i*th batch and summarize the overall variations in this batch with respect to the other batches over the entire history of the batch. The mixing matrix, $\mathbf{A}$, summarizes the time variations of the measured variables about their average trajectories. The elements of this matrix are the weights, which give the independent vectors $\mathbf{s}$ for a batch when applied to each variable at each time interval within that batch (Yoo et al., 2003).

Similar to MPCA, the key idea is to exploit the ability of MICA to extract features from three-way batch data by projecting the data onto a low-dimensional space that summarizes both the variables and their time trajectories. First, the three-way matrix $\underline{\mathbf{X}}(I \times J \times K)$ is unfolded into a two-dimensional matrix, $\mathbf{X}(I \times JK)$ using a batchwise unfolding scheme. Second, the mean trajectory is removed from each variable and each time of the unfolded data matrix to remove the majority of the nonlinear behavior of the batch process. Third, the data matrix is normalized (i.e., mean centered and standardized to unit variance). The normalized $\mathbf{X}(I \times JK)$ is then transposed, yielding the transposed matrix $\mathbf{X}_{normal}(JK \times I)$. Fourth, whitening is performed on $\mathbf{X}_{normal}(JK \times I)$ to acquire the uncorrelated whitened matrix $\mathbf{Z}_{normal} = \mathbf{Q}\mathbf{X}_{normal}$. Fifth, the matrices of $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{S}$ are obtained using the FastICA algorithm. Sixth, the procedures for ordering and dimension reduction method of ICs are executed. The $m$ rows of $\mathbf{W}$ constitute a reduced matrix $\mathbf{W_d}$ (deterministic part of $\mathbf{W}$), and the remainder of the rows of $\mathbf{W}$ constitute a reduced matrix $\mathbf{W_e}$ (excluded part of $\mathbf{W}$). Finally, the MICA model with the matrices $\mathbf{W_d}$, $\mathbf{W_e}$, $\mathbf{S_d}$ and $\mathbf{S_e}$ is constructed. Then, independent data vectors for a new batch $k$ ($\mathbf{x}_{new}(k)$), $\hat{\mathbf{s}}_{newd}(k)$ and $\hat{\mathbf{s}}_{newe}(k)$, can be obtained by transformation through the demixing matrices $\mathbf{W_d}$ and $\mathbf{W_e}$, i.e., $\hat{\mathbf{s}}_{newd}(k) = \mathbf{W}_d \mathbf{x}_{new}(k)$ and $\hat{\mathbf{s}}_{newe}(k) = \mathbf{W}_e \mathbf{x}_{new}(k)$, respectively.

In MICA, two statistics are deduced from the process model in normal operation: the *D*-statistic for the systematic part of the process variation and the *Q*-statistic for the residual part of the process variation. The *D*-statistic for a batch *k*, also known as the $I^2$ statistic, is the sum of the squared independent scores and is defined as follows:

$$I^2(k) = \hat{\mathbf{s}}_{newd}(k)^T \hat{\mathbf{s}}_{newd}(k) \tag{4}$$

The *Q*-statistic for a batch *k*, also known as the *SPE* statistic, is defined as follows:

$$SPE(k) = \mathbf{e}(k)^T \mathbf{e}(k) = (\mathbf{x}(k) - \hat{\mathbf{x}}(k))^T (\mathbf{x}(k) - \hat{\mathbf{x}}(k)) \tag{5}$$

where $\hat{\mathbf{x}}$ can be calculated as follows:

$$\hat{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{B}_d \hat{\mathbf{s}} = \mathbf{Q}^{-1}\mathbf{B}_d \mathbf{W}_d \mathbf{x} \tag{6}$$

The confidence limits of the $I^2$ and *SPE* statistics in MICA can be obtained by kernel density estimation. Here, the $I^2$ value is used to detect faults associated with abnormal variations within an MICA model subspace, whereas the *SPE* value is used to detect new events that are not taken into account in an MICA model subspace (Yoo *et al.*, 2003).

## 3. Result and Discussion

### 3.1 Process description of the pilot-scale SBR system

The data used in this research were collected from a pilot-scale SBR system shown in Fig. 1. A fill-and-draw sequencing batch reactor (SBR) with a 80-liter working volume is operated in a 6h cycle mode and each cycle consists of fill/anaerobic (1h), aerobic (2h 30 min), anoxic (1h), re-aerobic (30min) and settling/draw (1h) phases. The hydraulic retention time (HRT) and the solid retention time (SRT) are maintained at 12 hrs and 10 days, respectively. Six electrodes for pH, oxidation-reduction potential (ORP), dissolved oxygen (DO), temperature, conductivity and weight are connected to the individual sensors to check the status of the SBR, where a set of on-line measurements is obtained every one minute. The historical data set of the SBR consisted of 280 batches (70 days) for which 6 variables were measured at 300 time instants (Lee and Vanrolleghem, 2003).

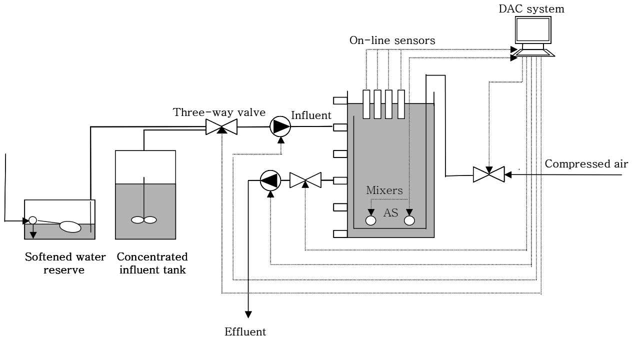

*Figure 1. Schematic diagram of the pilot-scale sequencing batch reactor.*

### 3.2 Multivariate analysis of historical data set in SBR (MPCA and MICA)

Fig. 2 shows the monitoring result of all 280 batches of the SBR using the MPCA and MICA methods, where the dotted lines correspond to the 95 and 99% confidence limits. Five components of the MPCA model were selected by the cross-validation method. To ensure comparison of equivalent models, five ICs were selected for the MICA model. From this figure, we notice that the MICA plot shows characteristics dissimilar from the MPCA one. Compared to MPCA, MICA points to a lower number of abnormal batches in SBR. This difference can be explained by the density estimation of the SBR data. Fig. 3 (left) shows that the density estimate of the first score ($t_1$) in MPCA does not follow the Gaussian distribution but the '*supergaussian distribution*' in which process variables take relatively more often values that are very close zero, where the probability density of the data is peaked in the middle and has heavy tails (large values far from zero). Thus, the $T^2$ and *SPE* charts of MPCA that are based on the assumption that the data are Gaussian distributed may cause a false result when it is used for SBR monitoring. This observation is the motivation of the MICA method because MICA is sensitive to modes whose influences on the measured variables follow a supergaussian distribution. Fig. 3(right) represents the loading plot of each variable of each time interval of the first IC. It shows the types of information that can be extracted when MICA is used in batch modeling. The loading plot obtained from MICA gives the history and identified important features of the SBR. From this figure, we notice that the DO, conductivity, and pH show large variations and have large influences during a batch, whereas ORP and weight show relatively small variations.

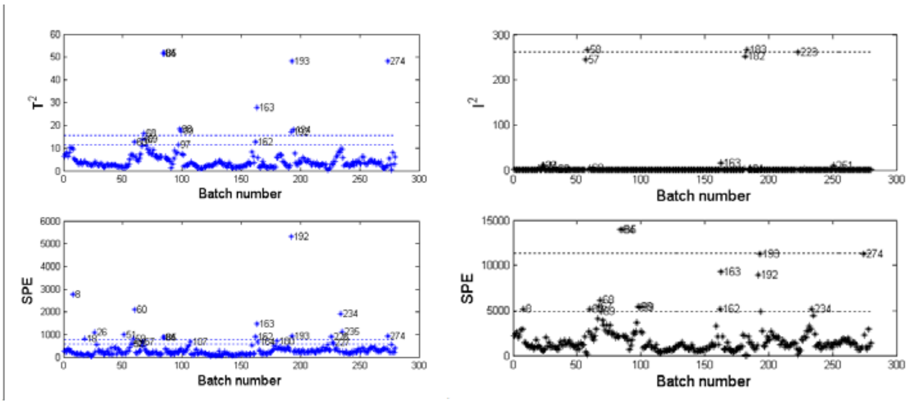

*Figure 2. Multivariate analysis of all 280 batches, (left) MPCA, (right) MICA.*

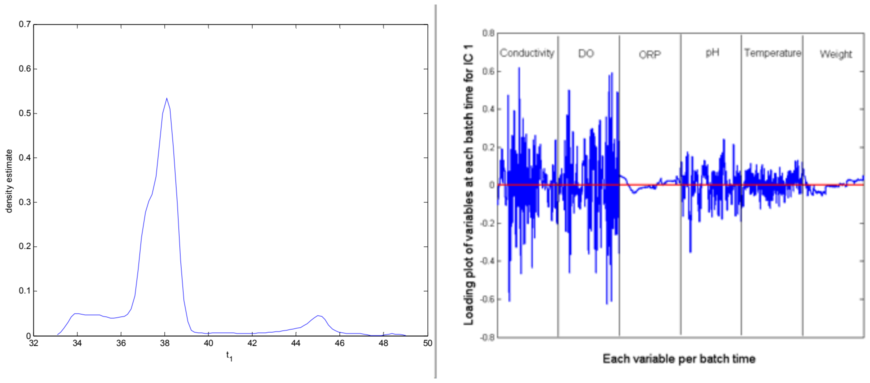

*Figure 3. The density estimate of MPCA and the variable loading plot of MICA. (left) Non-Gaussian distribution of the first principal score ($t_1$) obtained from MPCA, (right) Variable loading plot for the first independent score ($i_1$) obtained from MICA.*

### 3.3 Batch monitoring of SBR (MPCA and MICA)

The MPCA and MICA models for the SBR monitoring were developed after an analysis of the historical SBR data set in Fig. 2. The MPCA model selected 143 batches to create a rather broad scope of normal batches, where 7 abnormal batches (batch number: 8,18,26,51,60,84,85) were excluded for the normal operating condition (NOC) model. The MICA model selected 146 batches, where 4 abnormal batches (batch number: 57, 58, 84, 85) were excluded for the normal NOC model. The test data set that consisted of the following 30 batches was projected onto the reduced MPCA and MICA model spaces. Fig. 4 shows the batch monitoring result by MPCA and MICA. While both of them could detect two abnormal batches (batch 12, 13), MPCA detected batch 9 as an abnormal batch while MICA left batch 9 as a normal batch. Actually, batch 9 is a normal batch. When MPCA is applied to non-Gaussian data, the $T^2$ chart of MPCA may suffer oversensitivity for normal batches, e.g., batch 9. As a data set deviates from a Gaussian distribution, the variance tends to increase and hence the $T^2$ statistic tends to decrease. Typically, this increases the false alarm rate of the MPCA in which a normal batch might be judged as a non-conforming one. Obviously, this deteriorates the reliability of the monitoring system.

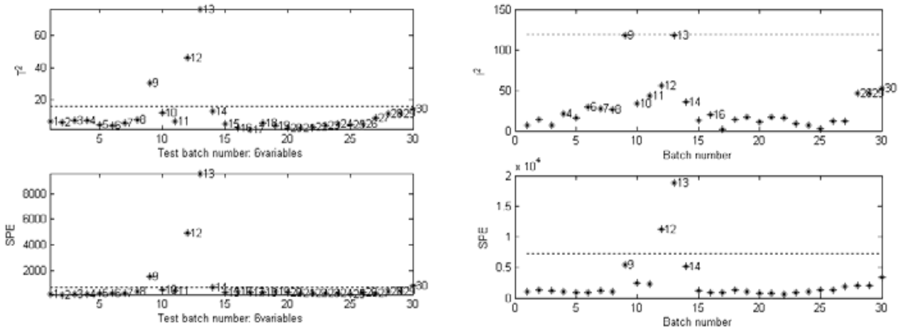

*Figure 4. Monitoring result of 30 test batches. (left) MPCA and (right) MICA. The dotted lines correspond to the 99% confidence limit.*

## 4. Conclusion

This paper describes the application of a pilot-scale SBR monitoring using MICA which can extract meaningful hidden information from non-Gaussian data sets. The result showed a more powerful monitoring performance than the MPCA approach. Furthermore, the MICA method can be easily applied to most batch or fed-batch processes which have non-Gaussian distributed data.

**Acknowledgement**

This work was supported by a Visiting Postdoctoral Fellowship of the Fund for Scientific Research-Flanders (FWO).

### References

Hyvärinen, A., J. Karhunen, and E. Oja, 2001, Independent component analysis, John Wiley & Sons, INC., USA.

Lee, D.S. and Vanrolleghem, P. A., 2003, Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis, Bio&Bioeng, 82, 489-497.

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

Yoo, C.K., J. Lee, P.A. Vanrolleghem and Lee, I.B., 2003, On-line monitoring of batch processes using multiway independent component analysis, Chemom. and Intel. Lab. Sys. (in revision).