Batch Process Monitoring by Three-way Data Analysis Approach

Ndèye Niang[1], Flavio S. Fogliatto[2], Gilbert Saporta[1]

[1]Chaire de Statistique Appliquée & CEDRIC CNAM
292, rue Saint Martin, 75141 Paris Cedex 03, France,
ndeye.niang_keita@cnam.fr
[2] Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil,
ffogliatto@producao.ufrgs.br

Abstract:
We propose a non parametric quality control strategy based on the three way method STATIS and convex hull peeling for monitoring batch processes with constant or variable duration. The method is illustrated on a simulated data set and on a real one.
Keywords: Multivariate quality control, STATIS method, Batch process.

1 Introduction

Batch processes are widely used in several industrial sectors, such as food and pharmaceutical manufacturing. In a typical batch, raw materials are loaded in the processing unit and submitted to a series of transformations, yielding the final product. Process performance is described by variables which are monitored as the batch progresses. Multivariate control charts are recommended for simultaneously monitoring several quality characteristics in a process or product. Traditional control charts are based on the independence and multinormality assumptions which are not always verifiable in practice; and are not efficient when the variables nominal behaviours are described by profiles. In those cases, process monitoring is usually accomplished using multivariate control charts (CCs) based on multiway principal components analysis (MPCA). These charts are denoted here by MPCA-CCs.

In this paper we propose a new quality control strategy for monitoring batch processes with constant and variable duration without any pre-treatment: the data set is reduced using the STATIS method (Lavit *et al.*, 1994) or its dual version. Two summarized representations of the batches become available. Monitoring of batch performance with respect to these two summarized representations above is accomplished directly on principal plane displays, from which non-parametric control charts based on convex hull peeling are derived. Our developments are built on the work of Scepi (2002), although offering original propositions for the on-line control of batches. Variable duration batch control was also not contemplated by that author.

Batch process control strategies have been proposed in the literature; some are presented next. Nomikos & MacGregor, 1995, Kourti & MacGregor, 1996 investigate the application of MPCA-CCs to monitor batch processes of fixed length by verifying the outputs of two control charts. The first is a T^2 chart for the scores obtained projecting future batches on the q principal components retained in the reference distribution derived running an MPCA on data coming from good batches that emerged from the process; it monitors the behavior of known process variability sources. The second is a Q chart for the residuals of the reference model; it detects any atypical events that disturb the correlation structure of the process variables. The Hotelling T^2control chart is based on a multinormality assumption and MPCA assume firstly that all batches considered in the analysis have the same length and are aligned with respect to process stages. When these two assumptions are not verified, MPCA-CCs must be adapted to handle variable batch duration and the use of non-parametric control charts should be considered.

Several approaches to the variable batch duration have been proposed (Kassidas *et al*., 1998). However, as briefly discussed below, propositions found in the literature are not always satisfactory. They globally consist of transforming the process variables to make them have the same length and then applying the MPCA-CCs. Authors such as Kassidas *et al*., 1998, and Doan & Srinivasan, 2008, proposed approaches where batches were aligned using dynamic time warping algorithms, which present some practical and theoretical intrinsic limitations; however, the greatest drawback in those approaches seems to be related to the representation of batch variation along the time axis, which is altered when stages in the batch process are synchronized.

Rosa, 2005, addressed the varying time batch problem using another approach based on the Hausdorff distance, which does not require dimension reduction techniques or procedures to align unequal batches. Despite its simplicity and the promising results obtained applying the method to simulated data sets, there is no evidence that the Hausdorff distance captures the original correlation structure.

2 A non parametric methodology

Our proposition is to consider each batch as a matrix where the rows are time instants. STATIS or Dual STATIS, outlined next, are then applied directly to the batch data; no pre-processing of batches with unequal duration is required. Finally a non parametric chart based on convex hull peeling is used to overcome the multinormality assumption.

2.1 STATIS

STATIS is a multivariate data analysis method aimed at exploring the three-way structure of several data tables obtained under different circumstances. The method was originally proposed by Escoufier, 2006, up to our knowledge its sole use in quality control is Scepi, 2002.

Let $\mathbf{X}_i$ (i = 1,…, N) be the i-th data matrix where T outcomes of p variables are available. Prior to analysis, matrices are usually normalized to remove scale effects in the original variables. STATIS associates to each $\mathbf{X}_i$ the ($T \times T$) matrix of scalar products $\mathbf{W}_i = \mathbf{X}_i\mathbf{X}_i'$, where $\mathbf{X}_i'$ denotes the transpose of $\mathbf{X}_i$. In order to compare two tables $\mathbf{X}_i$ and $\mathbf{X}_i$, STATIS uses the RV coefficient (for vector correlation), introduced by Robert and Escoufier, 2006, between $\mathbf{W}_i$ and $\mathbf{W}_{i'}$ defined as follows (considering two matrices):

$$RV_{ii'} = trace\left(\mathbf{W}_i\mathbf{W}_{i'}\right) \Big/ \sqrt{trace\left(\mathbf{W}_i\right)^2 trace\left(\mathbf{W}_{i'}\right)^2} \qquad (1)$$

RV is non negative and scaled between 0 and 1; the closer to 1, the more similar the matrices $\mathbf{W}_i$ and $\mathbf{W}_{i'}$. STATIS proceeds into two steps:

2.1.1 Inter *(between)* structure

Once coefficients between every pair of matrices are available, they are organized into a $\left(N \times N\right)$ square matrix **S**, which is diagonalized like in PCA. The similarities between the N data are visualized by projecting matrices $\mathbf{W}_i$ onto the first principal plane: the coordinate of batch i in the k-th factorial axis is given by $c_i^k = \sqrt{\lambda_k}u_i^k$ where λ_k is the eigenvalue associated with the k-th eigenvector $\mathbf{u}^k$ with entry u_i^k corresponding to the i batch.

2.1.2 Intra *(within)* structure

Since all elements of **S** are non negative, the first eigenvector $\mathbf{u}^1$ is a size factor. After standardization $\mathbf{u}^1$ gives $\boldsymbol{\alpha}^1$, whose entries α_i^1 are used to define what is called the compromise matrix **W** between the N data matrices:

$$\mathbf{W} = \sum\nolimits_{i=1}^{N} \alpha_i^1 \mathbf{W}_i \qquad (2)$$

The weights α_i^1 represent the agreement between data tables and the compromise. Eqn. (2) leads to a compromise matrix that is robust to outliers, which is desirable in quality control applications.

A second PCA performed now on **W** allows a visualization of artificial points B_t called compromise points, on the plane spanned by the first two principal components. Their coordinates in the k-th factorial axis are the elements of the following vector:

$$\mathbf{z}^k = \sqrt{\delta_k}\mathbf{v}^k = (1/\sqrt{\delta_k})\mathbf{W}\mathbf{v}^k \qquad (3)$$

where δ_k is the eigenvalue associated with the k-th eigenvector $\mathbf{v}^k$. Element z_t^k gives the coordinate of point B_t (t =1,…, T); this is the compromise plot.

Additionally, selecting points corresponding to a given observation t, we obtain a detailed representation of the variables joint behaviour in each data matrices for observation t. Such plot, denoted CO_t-graph enables an easy interpretation of changes in the N matrices for each observation. These additional studies constitute what is called the intrastructure analysis.

The compromise analysis described above is equivalent to another PCA on the unfolded matrix $\underline{\mathbf{X}}$ obtained by merging horizontally the data tables $\mathbf{X}_i$ weighted by α_i^1; i.e.:

$$\underline{\mathbf{X}} = \left[\sqrt{\alpha_1}\mathbf{X}_1 \quad | \sqrt{\alpha_2}\mathbf{X}_2 \quad | \quad \ldots \quad | \sqrt{\alpha_N}\mathbf{X}_N \right] \tag{4}$$

In all analyses above we assumed data matrices of same dimension. Note however that the analyses do not take into account column labels (only row labels), so STATIS may be used to analyze matrices having different number of columns but the same number of rows.

2.2 Dual STATIS

In cases where data matrices do not have the same number of rows, the method should be adapted which leads to "Dual STATIS". Let the data matrices $\mathbf{X}_i$ of dimension $(T_i \times p)$ be represented by their $(p \times p)$ covariance matrices $V_i = \mathbf{X}_i'\mathbf{X}_i$. The RV coefficients between pairs of matrices $\mathbf{V}_i$ and $\mathbf{V}_{i'}$ are defined in eqn. (5) and organized in a matrix $\mathbf{Z}$ with element $z_{ii'}$.

$$RV_{ii'} = z_{ii'} = trace\left(\mathbf{V}_i\mathbf{V}_{i'}\right) \Big/ \sqrt{trace\left(\mathbf{V}_i\right)^2 trace\left(\mathbf{V}_{i'}\right)^2} \tag{5}$$

The first two principal components, associated with the largest eigenvalues of $\mathbf{Z}$ obtained through PCA, give a graphical representation of the interstructure. The normalized elements β_i^1 of the first eigenvector of $\mathbf{Z}$ give the weights used to obtain the compromise matrix:

$$\mathbf{V} = \sum_{i=1}^{N} \beta_i^1 \mathbf{V}_i \tag{6}$$

Again, the first two principal components of $\mathbf{V}$ provide a graphical representation of the compromise and an intrastructure plot of the variables is obtained by projecting matrices $\mathbf{V}_i$ as supplementary points.

2.3 Non-parametric control charts

The non-parametric control charts (IS-CC and CO_t-CCs) we propose are based on convex hull peeling performed directly on principal plane displays obtained using the STATIS method: interstructure (IS) and intrastructure (CO_t) plots. To establish a $(1-\alpha)$ control region on such graphs we use the proposition in Zani *et al.* (1998), consisting into the following three steps. First, an inner region is defined in the plane such that a proportion π of the points in the graph fall within its boundaries. Such inner region is defined by a *B*-spline curve that smoothes the contours of a convex hull containing the points. π is equal to 50% of the points in Zani *et al.* (1998), but a larger proportion may also be used. Next, a robust centroid is determined as the mean of observations inside the inner region. Finally, the control region is established by defining a multiple of the distance l between the centroid and the boundary of the π-hull that corresponds to the desired probability α of false alarms. According to the authors, for α=0.01, the corresponding value of l is 1.68.

3 Applications

Off-line process control of a future batch *i= N+1* takes place initially by projecting X_{N+1} into the IS chart. When new batches have their coordinates falling inside the IS control region they are considered as in control. In case the projection yields an out-of-control signal, the CO_t charts are used to identify at which time instant the batch departed from the reference behaviour, or which variables are responsible for that.

3.1 Off-line control of batch process with same duration

One first example comes from of Nomikos and Mc Gregor widely used in the batch process literature (Eriksson et al). Data come from a batch polymerization reactor: 18 reference batches are selected to represent normal behaviour of the process. Additionally, a set of 11 batches is available to test the performance of the methods. This set contains 4 good batches and 7 bad batches. For each batch, 10 variables are recorded at 100 time instants. Variables X_1 X_2 X_3 X_6 and X_7 are temperature measurements variables X_4 X_8 and X_9 are pressure measurement, X_5 and X_{10} represent flow rates of materials added to the reactor. We apply STATIS and build the 99% control region on the summarized interstructure graph. Figure 1 shows the off line control results. All the 7 bad batches have been detected to be out-of control with stronger signal for 6 of them (fig.1.a). The out of control batch close to the boundaries was diagnosed as bad by the engineers and is known as having different behaviour and is generally not detected to be out of control (Eriksson et al). The use of CO_t charts (not shown for shortness) allows detection of departures from reference model at time instant 20 for the 6 batches strongly out of control. For good batches, fig.1.b displays the chart on which one of the 4 batches is signalled to be out of control. This is a false alarm already recorded in Eriksson et al. So, it can be said that our method performs better than or as well as classical method on this example.

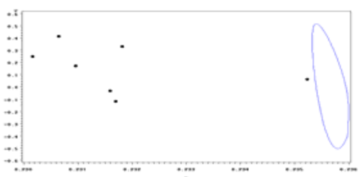

Fig. 1.a. Off line control - bad batches

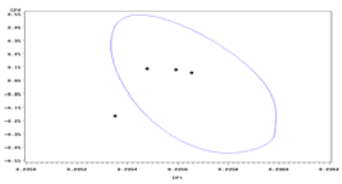

Fig. 1.b. Off line control - good batches

3.2 Off line control of batch process with variable duration

We exemplify the method proposed in section 2 using simulated data available in Rosa, 2005. The reference data set consists of 50 batches with duration varying from 89 to 100; three process variables are considered. Additionally, 2 good batches (GB1, GB2) and 2 bad one (BB1, BB2) are also considered in the analysis. All 3 variables have been perturbated to simulate bad batches. Fig.2.a and fig.2.b display the off line control results when applying DUAL STATIS: good batches projection fall within the control region while the bad batches are signalled to be out of control. The CO_t charts (not shown for shortness) signal all the 3 variables responsible for the out of control situation detected in fig.2.b.

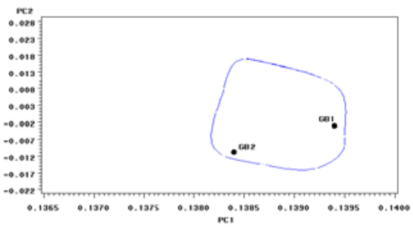

Fig. 2.a. Off line control - bad batches

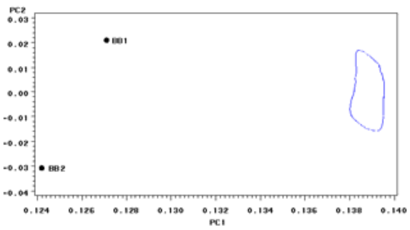

Fig. 2.b. Off line control - good batches

4 Conclusion

We have proposed a graphical non parametric strategy for quality control of batch processes of equal duration based on STATIS method for dimension reduction and its dual version in case of batches of varying duration. Process monitoring is accomplished through the use of two control charts. The IS control chart verifies batch progression compliance to the expected trajectories of its monitored variables. In the CO_t charts, trajectories of variables in a batch, are summarized and significant departures from their expected behaviour in the time axis are detected for batches with equal duration. More formal evaluations of the performances of the proposed strategies are necessary as well as their comparison to other methods in the literature. Extension to on line control of batch processes is in progress.

References

Doan, X-T., Srinivasan, R., 2008 Online monitoring of multi-phase batch processes using phase-based multivariate statistical control. *Computers and Chemical Engineering*, 32: 230-243.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S. 2001.*Multi- and Meag variate Data Analysis.* Umetrics

Escoufier, Y. 2006. Operator related to a data matrix: a survey. *Proceedings in Com putational Statistics* Rizzi A. *et al*. (eds), 285-297 Physica-Verlag.

Kassidas, A., MacGregor, J.F. and Taylor, 1998. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44: 864–875.

Kourti, T. and MacGregor, J.F. 1996. Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology*, 28 (4): 409–428.

Lavit, C. and Escoufier, Y., Sabatier, R. and Traissac, P. The ACT (STATIS method) *Computational Statistics & Data Analysis*, 18: 97-119.

Nomikos, P. and MacGregor, J.F. 1995. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1): 41–59.

Rosa, A.F.P. 2005 *Maîtrise Statistique de procédés par lots à temps variable*. Thèse de Doctorat de l'Université de Nantes, Nantes, France.

Scepi, G. 2002. Parametric and non parametric multivariate quality control charts. In *Multivariate Total Quality Control*, Physica-Verlag , Lauro C. *et al*. (eds), 163–189.

Zani, S., Riani, M. and Corbellini, A. 1998. Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28: 257-270.