ELSEVIER

Chemometrics and Intelligent Laboratory Systems 51 (2000) 95–114

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Generalized contribution plots in multivariate statistical process monitoring

Johan A. Westerhuis, Stephen P. Gurden, Age K. Smilde *

Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, Netherlands

Received 16 August 1999; accepted 13 February 2000

Abstract

This paper discusses contribution plots for both the *D*-statistic and the *Q*-statistic in multivariate statistical process control of batch processes. Contributions of process variables to the *D*-statistic are generalized to any type of latent variable model with or without orthogonality constraints. The calculation of contributions to the *Q*-statistic is discussed. Control limits for both types of contributions are introduced to show the relative importance of a contribution compared to the contributions of the corresponding process variables in the batches obtained under normal operating conditions. The contributions are introduced for off-line monitoring of batch processes, but can easily be extended to on-line monitoring and to continuous processes, as is shown in this paper.

Keywords: Contribution plots; Control limits; *D*-statistic; *Q*-statistic; Statistical process control

1. Introduction

Multivariate statistical process control (MSPC) using projection methods has been proven to be very useful for monitoring of industrial chemical processes [1,2]. Applications of MSPC for continuous processes as well as for batch processes have been very useful and work very well in practice [1–13]. The basis of this approach is to build an empirical model of a set of measurements obtained under normal operating conditions (NOC). Using this model, statistical confidence limits are calculated. New measurements are projected onto this model, and the statistics calculated should be within the confidence limits for the batch to be in control. The main problem with this approach is that in the case of a process disturbance, no information is obtained about the cause of the disturbance. This problem was already signaled by others [3,4,14,15] and one of the solutions to this problem was the use of contribution plots. A contribution plot shows the contribution of each process variable to the statistic calculated. A high contribution of a process variable usually indicates a problem with this specific variable. The use of contribution plots seems to work well in practice [3–7,15,16].

In the present paper, the theory of contribution plots is extended to latent variable models with correlated scores (e.g., PARAFAC, multiblock PCA or multiblock PLS). Furthermore, control limits for the contributions are introduced. These control limits help in finding the process variables that show dif-

* Corresponding author. Tel.: +31-20-525-5062; fax: +31-20-525-6638.

E-mail address: asmilde@its.chem.uva.nl (A.K. Smilde).

0169-7439/00/$ - see front matter
PII: S0169-7439(00)00062-9

ferent behavior compared to batches obtained under NOC. After a short introduction to the *D*-statistic and *Q*-statistic, which are used in MSPC, the contributions of each process variable to these statistics together with their corresponding control limits are presented. Different problems such as negative contributions to the *D*-statistic and smearing out of residuals over time and over different variables are discussed. A benchmark data set of a simulated semi-batch emulsion polymerization of styrene butadiene [17,18] is used to show the use of the contribution plots in practice.

2. Theory

The idea of plots of the contribution of each process variable to the *D*-statistic and *Q*-statistic is introduced for batch processes. At the end of this section, the ideas will be generalized to work for continuous processes as well.

2.1. Data

For batch processes, J process variables are measured for the whole batch duration at K specific time intervals. For batch i, this gives a matrix $\mathbf{X}_i(J \times K)$ of process measurements. For the statistical process control of such a process, I batches obtained under NOC are necessary to construct a statistical model of the process data. These I batches are assumed to capture all the common-cause variation present in the process. The measurements can be arranged in a three-way array $\underline{\mathbf{X}}(I \times J \times K)$. For convenience, in this paper, the three-way array of batch processes will always be considered matricized to a matrix $\mathbf{X}(I \times JK)$ where the batch direction is maintained [19]. Furthermore, $\mathbf{X}$ is always considered to be mean centered across the batch direction.

2.2. D- and Q-statistics

Using a set of I different batches obtained under NOC, an empirical latent variable model is developed to describe the data. The general form of this model equals:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

Here $\mathbf{X}$ contains the process data, $\mathbf{T}\mathbf{P}^{\mathrm{T}}$ is the model that contains the systematic part of the common-cause variation within the NOC data, and the residual matrix $\mathbf{E}$ contains the nonsystematic part not described by the model. $\mathbf{T}(I \times R)$ usually describes the difference between the batch runs, and $\mathbf{P}(JK \times R)$, which is the actual model, describes the similarities among the batch runs. The number of latent variables R is usually much smaller than I and JK. Several constraints for both $\mathbf{T}$ and $\mathbf{P}$ can be applied. For unfold principal component analysis, which is often used in MSPC to model $\mathbf{X}$, $\mathbf{T}$ is columnwise orthogonal and $\mathbf{P}$ is columnwise orthonormal, i.e. $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{D}$ and $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{I}$ is the identity matrix. For a PARAFAC model, $\mathbf{T}$ is not columnwise orthogonal and $\mathbf{P} = \mathbf{C} \odot \mathbf{B}$ where $\mathbf{C}(K \times R)$ and $\mathbf{B}(J \times R)$ are components in the time and process variable direction, respectively, and $\odot$ symbolizes the Khatri–Rao product [20].

From the process model, two types of statistics with known distributions are calculated. These are the *D*-statistic for the systematic part of the process variation and the *Q*-statistic for the residual part of the process variation. Using the distributions, confidence limits for the two statistics can be obtained. For the monitoring of new batches, the process data of the new batch $\mathbf{x}_{\mathrm{new}}(JK \times 1)$ is projected onto the model.

$$\mathbf{x}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{t}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} + \mathbf{e}_{\mathrm{new}}^{\mathrm{T}} \tag{2}$$

$$\mathbf{t}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)^{-1}$$

$$\mathbf{e}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}} - \mathbf{t}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}$$

In many models, $(\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}$ equals the identity matrix and thus $\mathbf{t}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}$.

The *D*- and *Q*-statistics calculated for the new batch run should be within the confidence limits for the batch to be in control. The *D*-statistic for the new batch, $\mathbf{x}_{\mathrm{new}}$, is defined as follows [21]:

$$D_{\mathrm{new}} = \mathbf{t}_{\mathrm{new}}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{t}_{\mathrm{new}} \sim \frac{R(I^2-1)}{I(I-R)}F(R, I-R, \alpha) \tag{3}$$

where $\mathbf{S}^{-1}$ equals the inverse of the covariance matrix of $\mathbf{T}$, $\mathbf{S}^{-1} = ((\mathbf{T}^{\mathrm{T}}\mathbf{T})/(I-1))^{-1}$. This *D*-statistic, divided by some constant, follows an *F*-distribution with R and $I-R$ degrees of freedom. The *Q*-

statistic for the nonsystematic part of the common cause variation of a new batch $\mathbf{x}_{\text{new}}$ is defined as follows:

$$Q_{\text{new}} = \sum_{jk=1}^{JK} \left(e_{\text{new},\,jk} \right)^2 \sim g\,\chi^2_{(h)} \tag{4}$$

where the scaling factor g and the degrees of freedom h are called the matching moments of the distribution. Nomikos and MacGregor [1] describe several ways to determine the confidence limits for the Q-statistic from the residuals $\mathbf{E}(I \times JK)$ of the batch runs obtained under NOC. In the present paper, the Jackson and Mudholkar approximation [22] is used. This approach uses a normal distribution to approximate the χ^2 distribution of the squared residuals.

$$Q_{\text{lim},\,\alpha} = \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1 - h_0}{\theta_1^2} \right) + \frac{\sqrt{z_\alpha \left(2\theta_2 h_0^2 \right)}}{\theta_1} \right]^{\frac{1}{h_0}} \tag{5}$$

where $h_0 = 1 - ((2\theta_1\theta_3)/(3\theta_2^2))$, $\theta_1 = tr(\mathbf{V})$, $\theta_2 = tr(\mathbf{V}^2)$ and $\theta_3 = tr(\mathbf{V}^3)$, $\mathbf{V}$ is the covariance matrix of $\mathbf{E}$, and z_α is the standardized normal variable with $(1 - \alpha)$ confidence limit, having the same sign as h_0.

For both the D- and Q-statistics, 95% or 99% confidence limits can be obtained. If the statistics of the new batch fall within these limits, the batch is considered to be in statistical control. The variation within this batch is considered as common cause variation and no special event took place during the batch.

2.3. Contribution plots

The main goal of MSPC of chemical processes is to detect variation in the process variables of a new batch that is different from common-cause variation in the process. However, the monitoring charts only detect that there is other variation in the process than the common-cause variation captured in the NOC data. This usually means that something is wrong with the process. The monitoring charts do not give information on what is wrong with the process, or which process variables caused the process to be out of control. By interrogating the underlying process model, at the point where an event has been detected, contribution plots may reveal the group of process variables making the highest contribution to the model or to the residuals [3,4,15,23].

In MSPC, the systematic variation and the residual variation are monitored with the D- and Q-statistics, respectively. In case of a process disturbance, one of these statistics or both will be above the confidence limits. If only the D-statistic is out of control, the model of the process is still valid, but the distance between the batch and the center of the model is too large. In this case, contributions of each process variable to the D-statistic should be examined. If the Q-statistic is above the confidence limits, a new event is found in the data, that is not described by the process model, presumably because this event was not present in the NOC data. In that case, the contributions of each process variable to the Q-statistic should be examined.

2.4. Contribution of the process variables to the Q-statistic

If, for a specific new batch, a disturbance was detected in the Q-chart of the residuals, then the contribution of the variables to the Q-statistic should be investigated. The contribution c_{jk}^Q of process variable j at time period k to the Q-statistic for this batch equals:

$$c_{jk}^Q = \left(e_{\text{new},\,jk} \right)^2 = \left(x_{\text{new},\,jk} - \hat{x}_{\text{new},\,jk} \right)^2 \tag{6}$$

where $x_{\text{new},\,jk}$ is the jkth element of $x_{\text{new}}(JK \times 1)$, $\hat{x}_{\text{new},\,jk}$ is the part of this element predicted by the model, and $e_{\text{new},\,jk}$ is the residual. In order to find at which time in the batch and for which variables the disturbance occurred, all contributions c_{jk}^Q can be plotted and examined. However, this approach may suffer from embedded error. This will be shown in the following simple example.

From an industrial batch process, a temperature from the heat source jacket (T_1), a temperature of the heat source coil (T_2) and the heat source supply pressure (P_1), which are highly correlated, were obtained. Measurements from 45 NOC batches were collected and after centering and scaling the data, a two-component unfold PCA model was build on this data. Fig. 1 shows the two loadings for each of the

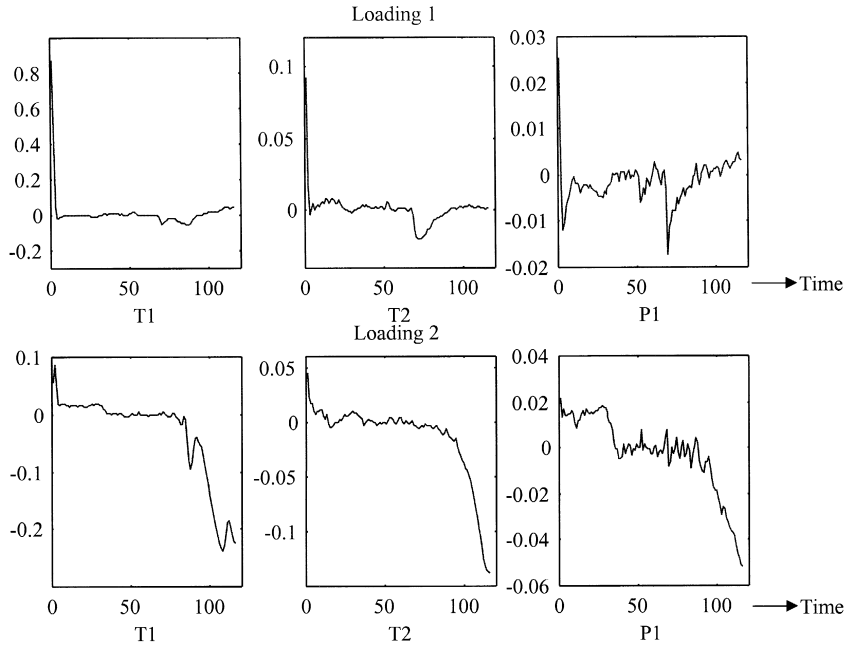

Fig. 1. First two loadings of example data set for temperatures T_1 and T_2 and pressure P_1.

temperatures and pressure that were calculated. Now, suppose a new batch $\mathbf{x}_{new}$ is obtained that for the first half of the batch run is perfectly in control, i.e. it behaves in exactly the same way as the mean of the 45 NOC batches. This mean is zero due to the centering procedure. Halfway through the batch, a sudden jump in temperature (T_1) and pressure (P_1) is measured, while the other sensor (T_2) stays on target. This batch is shown in Fig. 2. The new batch is projected on the model and scores, residuals and $\hat{x}_{new}$ can be calculated. Fig. 3 shows $\hat{x}_{new}$ and the residuals $e_{new,\,k}$ for both temperatures and pressure of the new batch. A clear jump is observed for the residuals of temperatures T_1 and pressure P_1 halfway through the reaction. However, for temperature T_2 and also at the beginning of the batch for temperature T_1 and pressure P_1, nonzero residuals are observed where the variables were still in perfect control, and where the residuals were expected to be zero. With the projection of the new data onto the model, all the measurements are compressed into only two score values. For the calculation of the residuals, the information is ex-

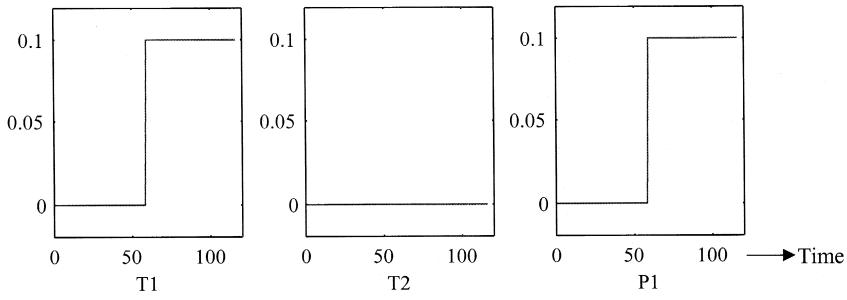

Fig. 2. Trajectories for temperatures T_1, T_2 and pressure P_1 for simulated new batch.

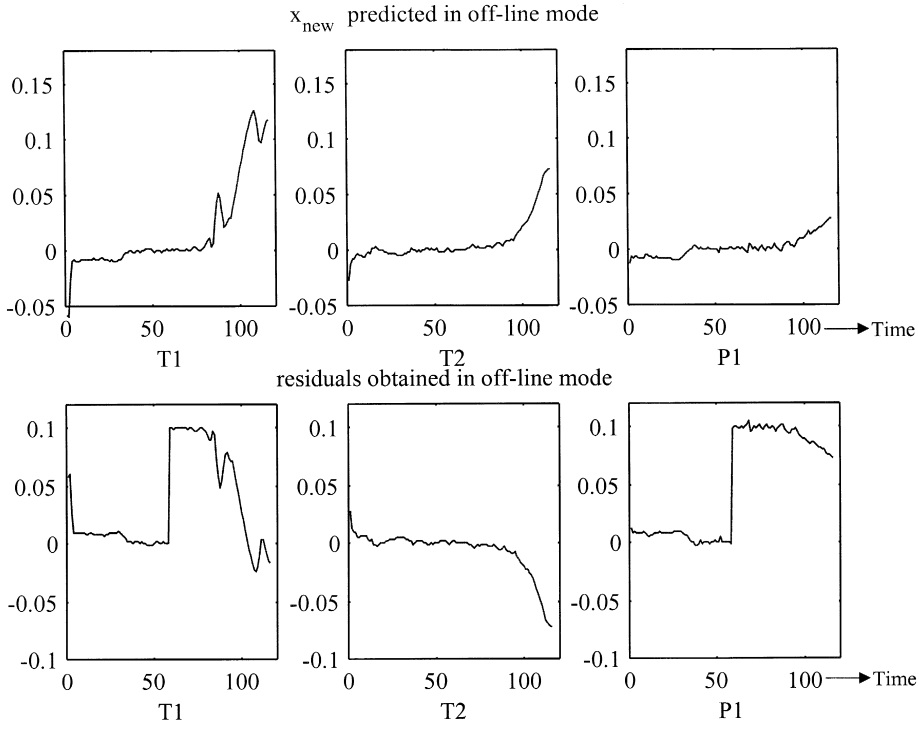

Fig. 3. Predicted new batch and residuals for T_1, T_2 and P_1 using off-line approach.

tracted from the two score values. Due to this compression and extraction, the information is spread out over the whole batch. This leads to the nonzero residuals in the batch, at positions where they were expected to be zero.

A partial cure to this smearing out of information is to consider the residuals as if they were obtained in an on-line mode. This means that at time k in the batch only the scores and residuals at time k are determined. The same approach as the on-line strategy for process monitoring presented by Nomikos and MacGregor [1] can be used. This means that with each new measurement coming in, the remainder of the batch is filled in using, e.g., the current deviations approach, then the data is projected on the model and scores and residuals can be calculated. The current deviations approach assumes that for each of the process variables, the deviation from the mean will stay the same for the remainder of the batch. This deviation value is then filled in for the remainder of the matrix with process measurements. Using this approach, the results of $\hat{x}_{\text{new}}$ and the residuals shown in Fig. 4 are obtained. Now the residuals at the beginning of the batch are zero as expected. Thus, using the on-line approach prevents the smearing out of residuals in the time direction.

Another very important issue of using contribution plots for residuals is also visible in Figs. 3 and 4. Although, temperature T_2 follows exactly the average batch trajectory, the residuals become large at the end of the batch. However, halfway through the batch, at time 60, residuals of T_1 and P_1 were already high. This means that from that point on, the model is not valid anymore. Therefore, any scores and residuals later on in the batch cannot be trusted.

Concluding, if the Q-statistic is outside control limits, residuals should be examined to find the time and process variables that caused the error. However, due to embedded error, residuals are smeared out over time and over the different process variables. The smearing out of residuals in the time direction can be solved by calculating residuals as if they were obtained in an on-line mode. The smearing out of residuals over different process variables cannot be solved in this way. After the first high residuals have been detected, the model is not valid anymore and the

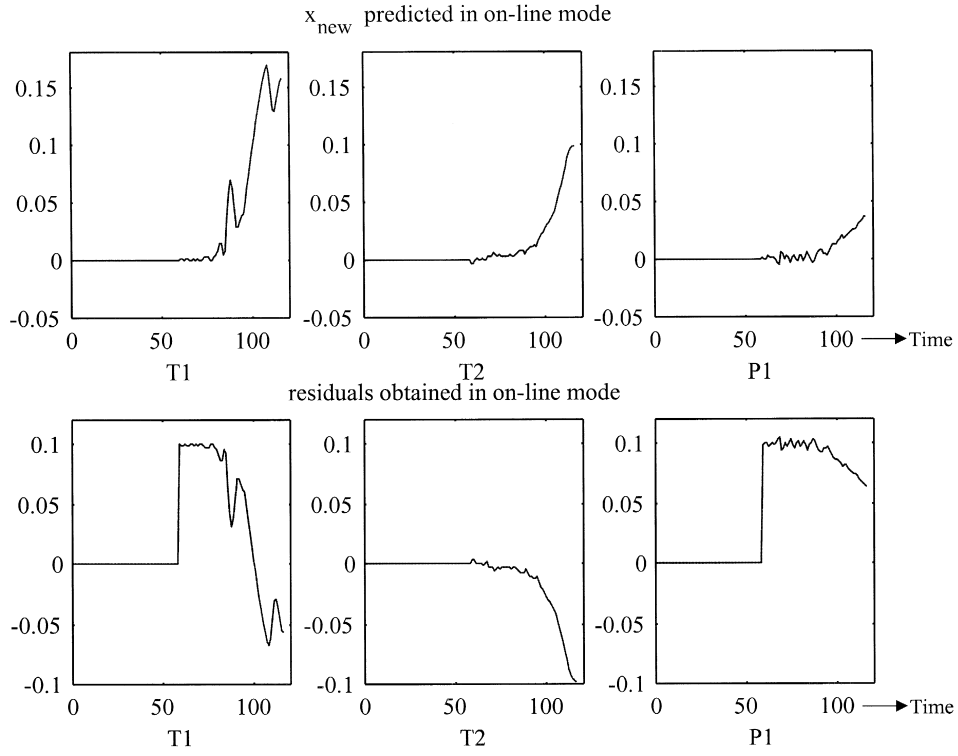

Fig. 4. Predicted new batch and residuals for T_1, T_2 and P_1 using on-line approach.

residuals for the remainder of the batch cannot be trusted. The high residuals found in this way should be considered, and using engineering knowledge from the plant, the problem should be addressed. Note that process variables that are in control can also give high residuals due to a mismatch of the model.

When the number of process variables is large, it makes sense to sum the residuals of each process variable for each time period, in order to detect at which point during the batch the disturbance took place.

$$c_k^Q = \sum_{j=1}^{J} c_{jk}^Q = \sum_{j=1}^{J} \left(e_{\text{new},jk} \right)^2$$

$$= \sum_{j=1}^{J} \left(x_{\text{new},jk} - \hat{x}_{\text{new},jk} \right)^2 \qquad (7)$$

Here, $e_{\text{new},jk}$ is the residual of the new batch of process variable j at time k, obtained using the on-line approach described above. It is usually easy to find the periods at which the disturbances took place. Then the contribution of all process variables to the summed residuals can be examined to find those variables that caused the process disturbance.

If on-line monitoring of the batch process is used, it is immediately known at which time period the special event took place in the process. One can directly zoom in on this period to find the process variables that are responsible for this special event. Applications of on-line monitoring of batch processes with contribution plots of the residuals can be found in Refs. [6,23]. However, the smearing out effect of residuals is not dealt with in those applications.

After the process variables that caused the disturbance of the process have been detected, it is recommendable to show the trajectory of the process variable that caused the disturbance together with its normal trajectory (which is the mean of the trajectories of all NOC batches for this specific process variable) as a confirmatory practice following a diagno-

sis of contribution plots. This seems to be an important psychological reinforcement to engineers and operators [3]. It also shows whether the specific process variable was higher or lower than usual during the special event. Furthermore, it can give information on the type of disturbance detected, e.g., a slow drift or a fast level change.

2.4.1. Control limits for contributions to Q-statistic

Plots of the contribution to the *Q*-statistic are similar to standard plots of squared residuals obtained in an on-line mode, but the contribution plots presented here also have control limits. These control limits are used to compare the residuals of the new batch to the residuals of the NOC data. If in the NOC data a certain process variable had high residuals, it can also be expected to have high residuals in the new batch. However, if a new batch has high residuals for a certain process variable that had low residuals in the NOC data, this probably is due to a special event in the new batch. Thus, instead of considering the absolute size of the residuals, the relative size, compared to the NOC residuals, should be examined. If the contributions of a large group of process variables are studied, it is usually found that several process variables have high contributions. Using the control limits, it is easier to find those process variables that are really different, compared to the NOC data. The control limits are calculated in the same way as the *Q*-statistic confidence limit, i.e. using the Jackson and Mudholkar approximation as presented in Eq. (5). For every combination of process variables or time periods, the control limits can be calculated according to this approximation when the corresponding columns of the NOC residuals **E** are used. However, for the contribution plots, the residuals that are obtained in an on-line mode should be used, as was described above. Every subset of columns of the on-line **E** still follows the χ^2 distribution, but the matching moments g and h will be different.

2.5. Contributions of process variables to the D-statistic

If only the *D*-statistic is out of control, the model is still valid, but the scores of the new batch have a larger Mahalanobis distance to the center of the model than the batches obtained under NOC. To find the process variable that caused the scores to be different, contributions of each process variable to the scores can be determined. Conceptually, contributions are different from loadings. Loadings represent variability across the entire NOC data set. Contributions represent the particular process variables that were unusual for a new given batch [3]. Two different approaches for calculating contributions were introduced. The first type of contributions, introduced by Miller et al. [3] and by MacGregor et al. [4] is the contribution of each process variable to a separate score. The first step in this approach is to find the specific **t** score that is above its own confidence limits. The confidence limits for a separate score are defined as follows:

$$0 \pm t_{\left(I-1,\frac{\alpha}{2}\right)} s_r \sqrt{\left(1+\frac{1}{I}\right)} \tag{8}$$

where a t-distribution with $I-1$ degrees of freedom is assumed, and s_r is the standard deviation of the rth score vector. The scores are assumed to be normally distributed with mean zero. Nomikos and MacGregor [1] conclude that the scores are well approximated by a normal distribution except for the first few time periods of the batch. The second step of this approach is to calculate the contribution of each element of the new batch run $x_{\text{new},jk}$ on the rth score for a PCA model [3]:

$$c^{t_r}_{jk,r} = x_{\text{new},jk} p_{jk,r} \tag{9}$$

Here, the summed contributions equal the $t_{\text{new},r}$ score of the new batch. If more scores are outside confidence limits, which is often the case, contributions have to be summed over these scores. This approach assumes that **P** of the model is columnwise orthonormal. This is the case in PCA models, but this is not general for all models. Furthermore, the approaches assume orthogonal scores. If the scores are not orthogonal, it is possible that the *D*-statistic is above its confidence limits, but that none of the scores is above its own confidence limits. This is caused by the fact that the correlation between the scores is neglected if univariate control is used on separate scores. Thus, if the scores are not orthogonal, the approach of contributions to separate scores cannot be used in general.

The second approach of contributions to the *D*-statistic was introduced by Nomikos [15]. This ap-

proach calculates contributions of each process variable to the *D*-statistic instead of to the separate scores.

$$c_{jk}^{D} = \sum_{r=1}^{R} \mathbf{S}_{rr}^{-1} t_{\text{new},r}\, x_{\text{new},jk}\, p_{r,jk} \qquad (10)$$

Here, the contribution of each element $x_{\text{new},jk}$ to the *D*-statistic is calculated. This contribution is summed over all r components. The summation can be used in case of orthogonal scores, because then $\mathbf{S}^{-1}$ is diagonal and only the diagonal elements of $\mathbf{S}^{-1}$ need to be considered. Furthermore, the loadings $\mathbf{P}$ of the process model are also assumed to be orthogonal, $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$.

Although many statistical data analysis methods will give orthogonal scores and loadings, this is not generally the case. Some examples of nonorthogonal scores can be found in literature. The first example is the PLS method defined by Martens and Naes [25], which gives the same predictions as standard PLS, but does not have the restriction of orthogonal scores. The second example is the use of multiblock PLS where the block scores are used for monitoring. If the super score deflation method [26] is used, then the block scores are nonorthogonal. MacGregor et al. [4] showed the use of multiblock PLS for the monitoring of an LDPE process; however, in that approach, block score deflation MBPLS was used, which leads to orthogonal block scores. Recently, three approaches of using PARAFAC models for the monitoring of batch processes were presented. Dahl et al. [8] compared the PARAFAC model with an unfold PCA model and Boqué and Smilde [7] compared a PARAFAC structure to a Tucker3 structure in a multiway covariates regression model. Louwerse and Smilde [27] compare unfold PCA, PARAFAC and Tucker3 structures to model three-way batch data. PARAFAC loadings in general are not orthogonal and also cannot be rotated to orthogonality without changing the solution. Louwerse and Smilde [27] also used a corrected way of calculating the score values of each NOC batch by projecting this batch on a model developed from all other NOC batches. This corrected version also leads to nonorthogonal $\mathbf{T}$ and nondiagonal $\mathbf{S}$.

According to Nomikos [15], the approach presented above for the calculation of the contributions is an approximation and valid only for principal

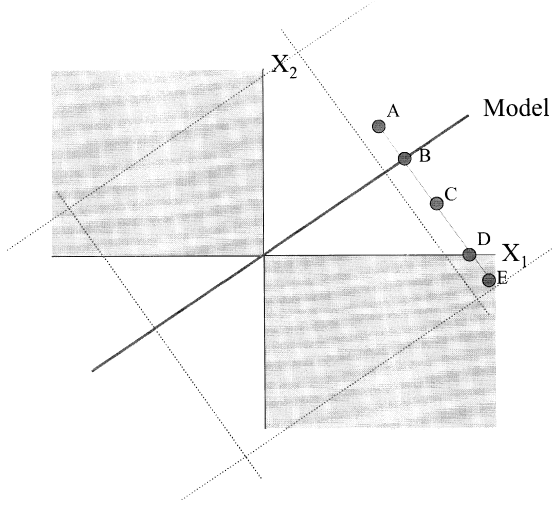

Fig. 5. Simple one-component model of two process variables. The dotted lines represent the confidence limits for *D*-statistic (perpendicular to model) and *Q*-statistic (parallel to model). Negative contributions are encountered only in the gray parts. Measurements A–E, on a line orthogonal to the model are explained in the text.

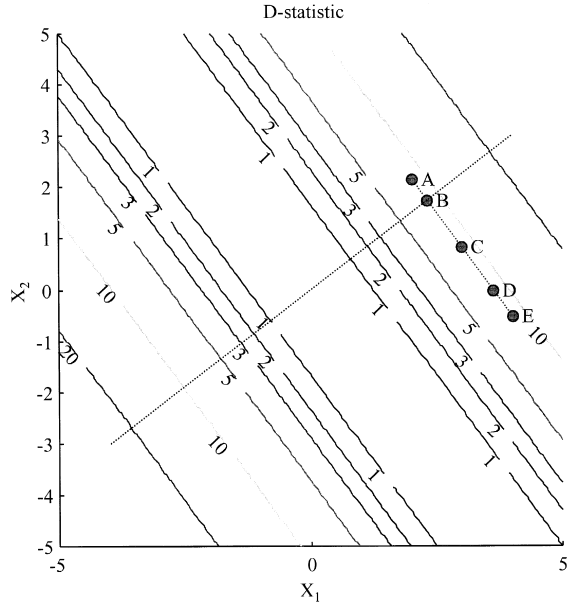

Fig. 6. *D*-statistic for model presented in Fig. 5. The dotted line represents the model. Measurements A–E, on a line orthogonal to the model are explained in the text.

component decompositions. The approach presented in the present paper is a generalization of the work of Nomikos [15] to also work for nonorthogonal scores and loadings. The *D*-statistic for a new batch $\mathbf{x}_{new}(JK \times 1)$ is defined as:

$$D_{new} = \mathbf{t}_{new}^{T}\mathbf{S}^{-1}\mathbf{t}_{new} = \mathbf{t}_{new}^{T}\mathbf{S}^{-1}\left[\mathbf{x}_{new}^{T}\mathbf{P}(\mathbf{P}^{T}\mathbf{P})^{-1}\right]^{T}$$
$$= \mathbf{t}_{new}^{T}\mathbf{S}^{-1}\sum_{jk=1}^{JK}\left[x_{new,jk}\mathbf{p}_{jk}^{T}(\mathbf{P}^{T}\mathbf{P})^{-1}\right]^{T}$$
$$= \sum_{jk=1}^{JK}\mathbf{t}_{new}^{T}\mathbf{S}^{-1}\left[x_{new,jk}\mathbf{p}_{jk}^{T}(\mathbf{P}^{T}\mathbf{P})^{-1}\right]^{T}$$
$$= \sum_{jk=1}^{JK} c_{jk}^{D} \qquad (11)$$

Thus, the contribution of element $x_{new,jk}$ of the new batch $\mathbf{x}_{new}$ to the *D*-statistic equals:

$$c_{jk}^{D} = \mathbf{t}_{new}^{T}\mathbf{S}^{-1}\left[x_{new,jk}\mathbf{p}_{jk}^{T}(\mathbf{P}^{T}\mathbf{P})^{-1}\right]^{T} \qquad (12)$$

Here, $\mathbf{t}_{new}^{T} = \mathbf{x}_{new}^{T}\mathbf{P}(\mathbf{P}^{T}\mathbf{P})^{-1}$ and $\mathbf{S}^{-1}$ is the inverse of the covariance matrix of the scores **T** of the NOC model. This approach is general and can be used for any type of model without restrictions to **T** and **P**. This generalization leads to Eq. (10) for models with $\mathbf{T}^{T}\mathbf{T} = \mathbf{D}$ and $\mathbf{P}^{T}\mathbf{P} = \mathbf{I}$.

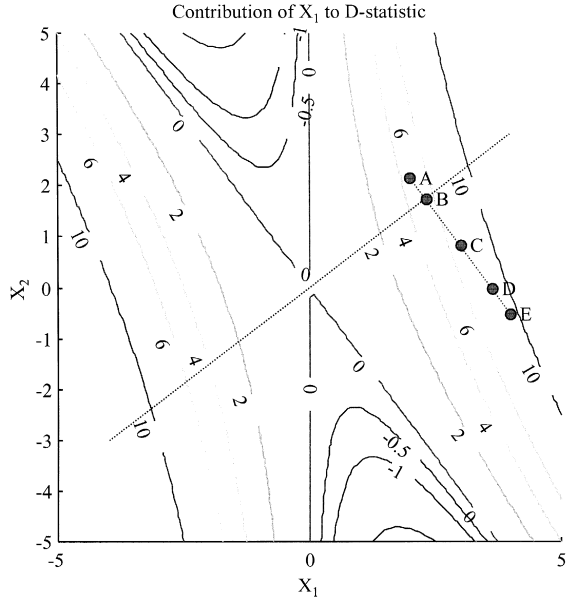

Fig. 7. Contributions of process variable X_1 to *D*-statistic of model presented in Fig. 5. The dotted line represents the model. Measurements A–E, on a line orthogonal to the model are explained in the text.

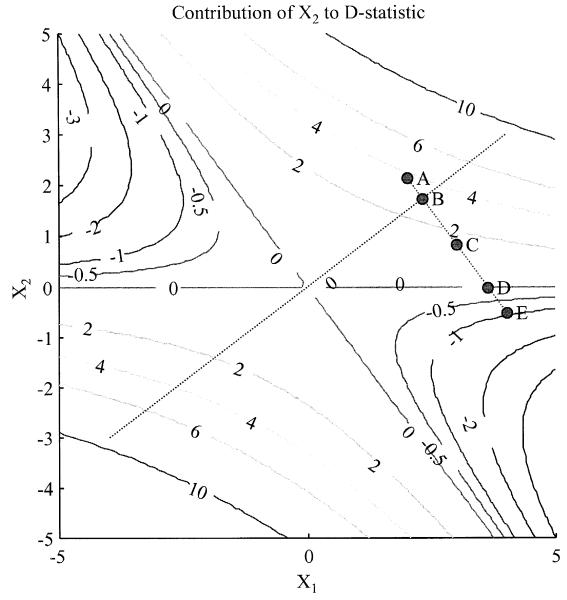

Fig. 8. Contributions of process variable X_2 to *D*-statistic of model presented in Fig. 5. The dotted line represents the model. Measurements A–E, on a line orthogonal to the model are explained in the text.

2.5.1. Negative contributions

The contribution of process variable $x_{new,jk}$ to the *D*-statistic, as presented in Eq. (12), can be positive or negative, although the sum of all contributions is non-negative because that equals the *D*-statistic.

Table 1
Process variables obtained from a simulated batch process of emulsion polymerization of styrene butadiene rubber

1	Flow rate of styrene
2	Flow rate of butadiene
3	Temperature of the feed
4	Temperature of the reactor
5	Temperature of the cooling water
6	Temperature of the reactor jacket
7	Density of the latex in the reactor
8	Estimate of total conversion
9	Estimate of instantaneous rate of energy release

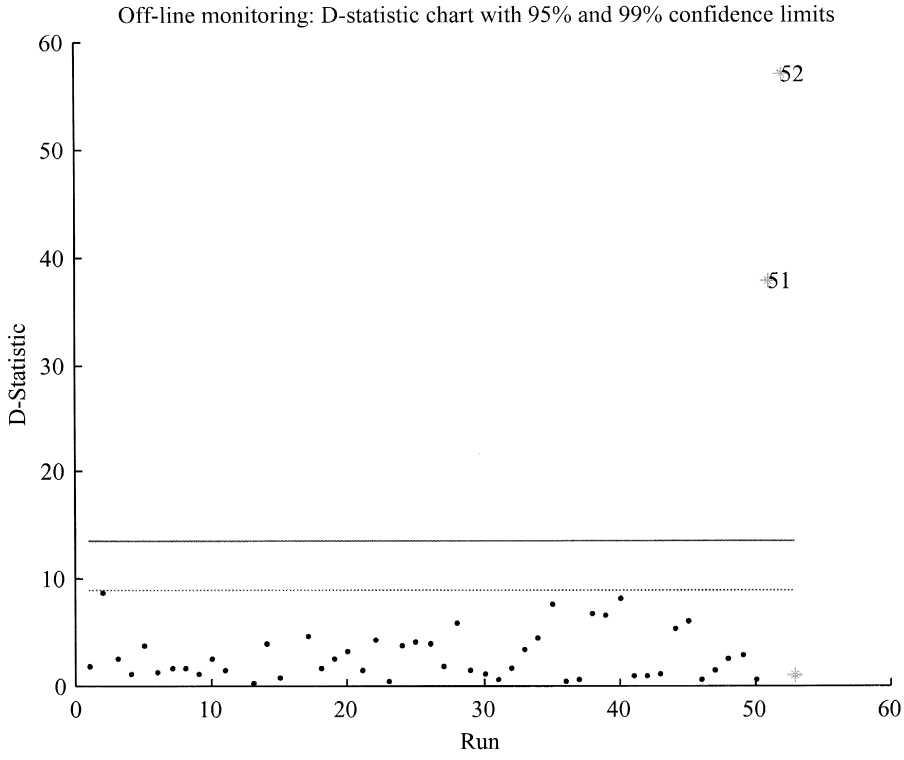

Fig. 9. *D*-statistic chart with 95% and 99% confidence limits. The dots represent the NOC batches and the stars represent the new batches.

Negative contributions were already encountered earlier by Kourti and MacGregor [23]. For a better understanding of negative contributions, the **t**-score of Eq. (12) is expressed as a summation of its separate multiplications of the data x_a times the loading $\mathbf{p}_a$.

$$c_{jk}^{D} = \sum_{a=1}^{JK} x_a \mathbf{p}_a^{\mathrm{T}} (\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1} \mathbf{S}^{-1} \times \left[x_{jk} \mathbf{p}_{jk}^{\mathrm{T}} (\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1} \right]^{\mathrm{T}} \quad (13)$$

If only one latent component is calculated, $(\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}$ and $\mathbf{S}^{-1}$ are just scaling constants and Eq. (13) can be written, neglecting these constants, as:

$$c_{jk}^{D} = \sum_{a=1}^{JK} x_a p_a p_{jk} x_{jk} = x_1 p_1 p_{jk} x_{jk} + x_2 p_2 p_{jk} x_{jk} + \ldots + x_{JK} p_{JK} p_{jk} x_{jk} \quad (14)$$

Eq. (14) shows that the contribution of process variable x_{jk} to the *D*-statistic is a summation of the separate multiplications of $x_{jk} p_{jk}$ times $x_a p_a$ of all other process variables. Each of these parts can be positive or negative except for the part where $a = jk$, which equals $(x_{jk} p_{jk})^2$ and is non-negative.

To study the negative contributions, a simple example of a one-component model with two process variables is presented. The loadings of the model are 0.8 and 0.6 for X_1 and X_2, respectively. Fig. 5 shows this model represented by the bold diagonal line. The score value of each combination of X_1 and X_2, equals the projection on the model. The *D*-statistic is the square of the score value and the distance perpendicular to the model is represented by the *Q*-statistic. The *D*-statistic increases when moving away from the origin in the direction of the model, and the *Q*-statistic increases in the direction orthogonal to the model. The confidence limits for the *D*-statistic are the dotted lines perpendicular to the model and the

confidence limits for the *Q*-statistic are the dotted lines parallel to the model. Inside the dotted box, the measurements are considered in control. The contribution to the *D*-statistic in this example for X_1 equals:

$$c_{x1}^{D} = (x_1 p_{x1})^2 + x_1 p_{x1} x_2 p_{x2} \quad (15)$$

The contribution of process variable X_2 is calculated in the same way as the contribution for X_1. The sum of these two contributions equals the *D*-statistic (see Fig. 6). Figs. 7 and 8 show the contribution of process variable X_1 and X_2 on the *D*-statistic for each combination of X_1 and X_2. The contribution of X_1 is zero for $X_1 = 0$ and for the combination of X_1 and X_2 on the line that projects onto the origin of the model. Between these two lines, c_{x1}^{D} is negative. The contribution of X_2 to the *D*-statistic is zero for $X_2 = 0$ and for combination of X_1 and X_2 on the line that projects onto the origin of the model. Between these two lines, c_{x2}^{D} is negative. Thus, only in the gray parts of Fig. 5, negative contributions are observed for either one of the variables X_1 or X_2.

Consider the measurements *A*–*E* shown in Fig. 5, represented by the circles. All these measurements are on the same line orthogonal to the model, and thus project to the same position on the model (*B*). The *D*-statistic, which is equal for these measurements equals 8.42. At the model line in measurement *B*, the contribution for X_1 is 5.38 and the contribution for X_2 is 3.02. The ratio between the two contributions c_{x1}^{D} and c_{x2}^{D} on the model line always equals the ratio between the squared loading values, $(0.8)^2/(0.6)^2 = 1.78$. If, along the line orthogonal to the model the value of X_2 is increased towards measurement *A*, then c_{x2}^{D} increases and c_{x1}^{D} decreases. In that case, the high value of X_2 is increasingly important for the *D*-statistic to be high. If, along the same line X_2 is decreased towards measurement *C*, then c_{x2}^{D} decreases and c_{x1}^{D} increases. In that case, the high value of the *D*-statistic is mainly caused by a large X_1 value. If X_2 is further decreased along the line, c_{x2}^{D}

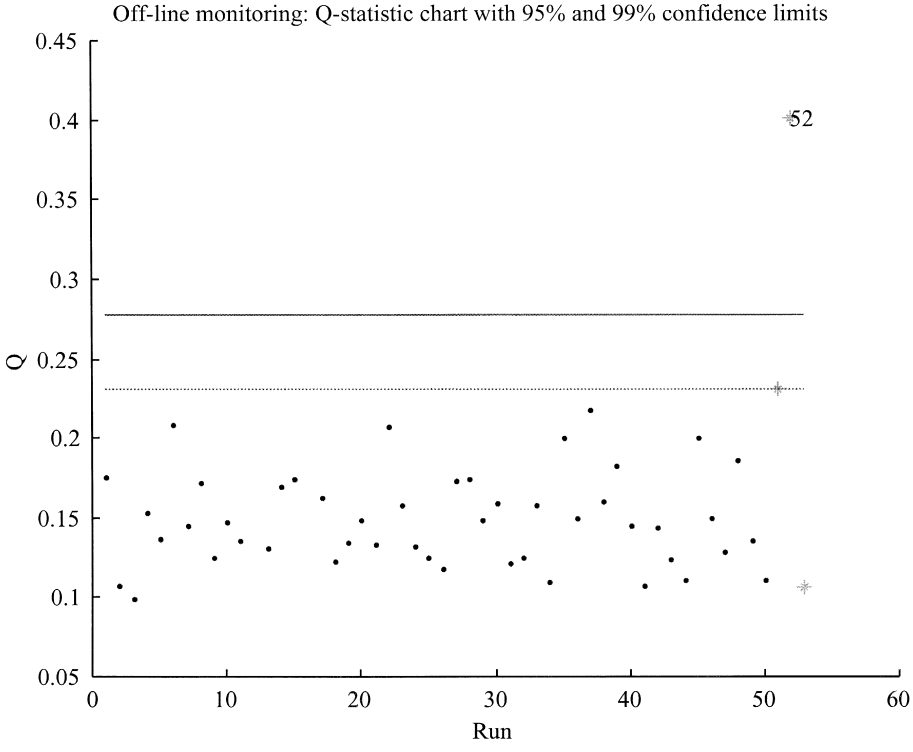

Fig. 10. *Q*-statistic chart with 95% and 99% confidence limits. The dots represent the NOC batches and the stars represent the new batches.

will keep decreasing, until it becomes zero for $X_2 = 0$ in measurement D, and even negative for $X_2 < 0$ in measurement E. This shows that a negative contribution is not a special event. It only forces the other contributions to be even higher.

Negative contributions are obtained when the signs of two process measurements are different than expected from the signs of the corresponding loadings. An object with a negative contribution may sometimes also have a Q-statistic that is outside the confidence limit. In that case, the Q chart will detect the disturbance. An object such as E, might be caused by a large value of X_1, but on the other hand, the value of X_2 might also be off. However, in the latter case, a value for X_2 which is more expected than the one measured, for the corresponding value of X_1, would lead to a D-statistic that is even higher than the one obtained. Therefore, a high contribution for X_1 in this case is a reasonable conclusion.

2.5.2. Control limits of D-contribution plots

Just as in the contribution plots of the Q-statistic, control limits to the D-statistic help to find the process variables that are different in this new batch compared to the NOC batches. However, in this case, it is not possible to use the same F-distribution as used for the D-statistic. The control limits are therefore obtained using a jackknife procedure in which each of the NOC batches is left out once, and contributions are calculated for each of the process variables of the batches left out.

The mean and variance of all I contributions of each jth process variable at the kth time period can be determined and will be used to obtain control lim-

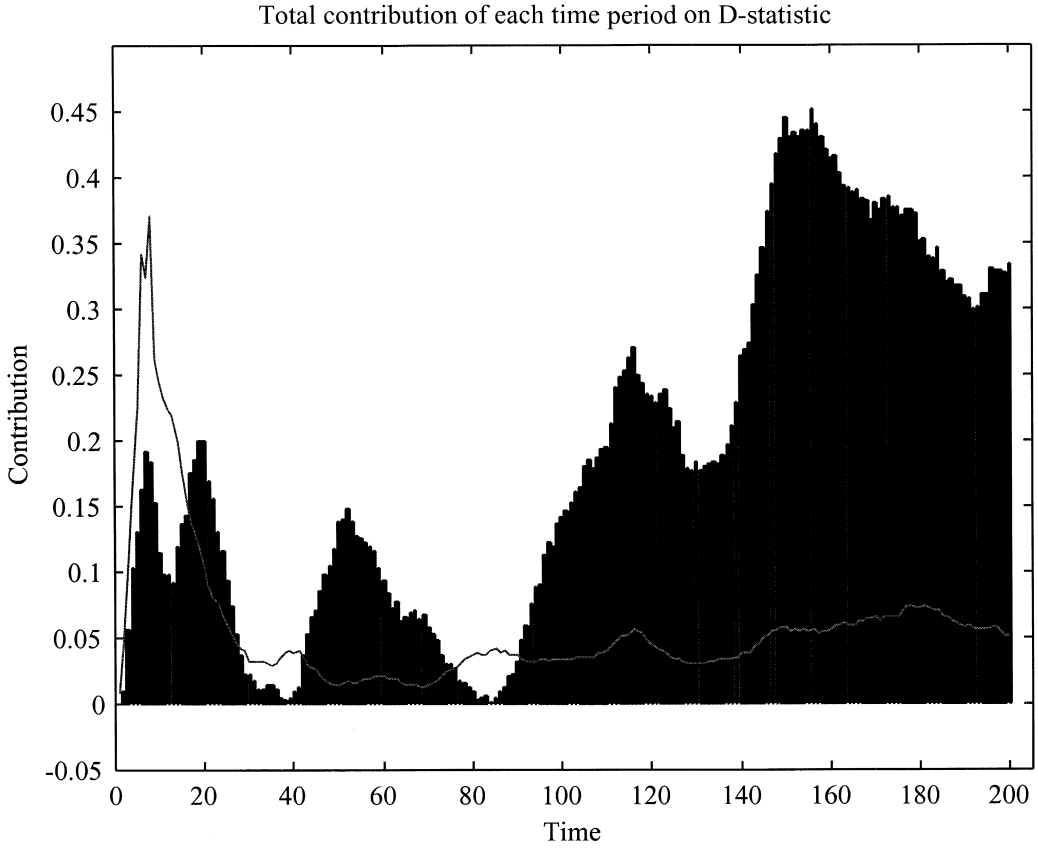

Fig. 11. Total contributions to the D-statistic of batch 51 of all process variables summed for each time period. Control limits represent the contributions expected from NOC batches. This batch is known to have a disturbance from the start of the process.

its for contributions of new batches. The upper control limit (UCL) for the contribution for each process variable is calculated as the mean of the contributions plus three times the standard deviation of the contributions for each process variable at each time period. The lower control limit is not used because only high contributions force the *D*-statistic to be out of control.

These UCL must not be considered to have statistical significance, but are very helpful in detecting contributions that are higher than contributions of NOC batches. If the contributions are summed over all process variables or over all time periods, then the UCL is obtained by summing the means of the corresponding jackknifed contributions. The standard deviation of this summed mean equals the square root of the summed squared standard deviations of the corresponding jackknifed contributions

$$s_k = \sqrt{\sum_{j=1}^{J} s_{jk}^2} \tag{16}$$

2.6. Continuous processes

In continuous processes, J process variables are measured. For the statistical process control of a continuous process, a set of I measurements obtained under NOC $\mathbf{X}(I \times J)$ is used to construct an empirical process model.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{17}$$

Again $\mathbf{X}$ contains the process data, $\mathbf{P}(J \times R)$ is the model, $\mathbf{T}(I \times R)$ describes the differences between

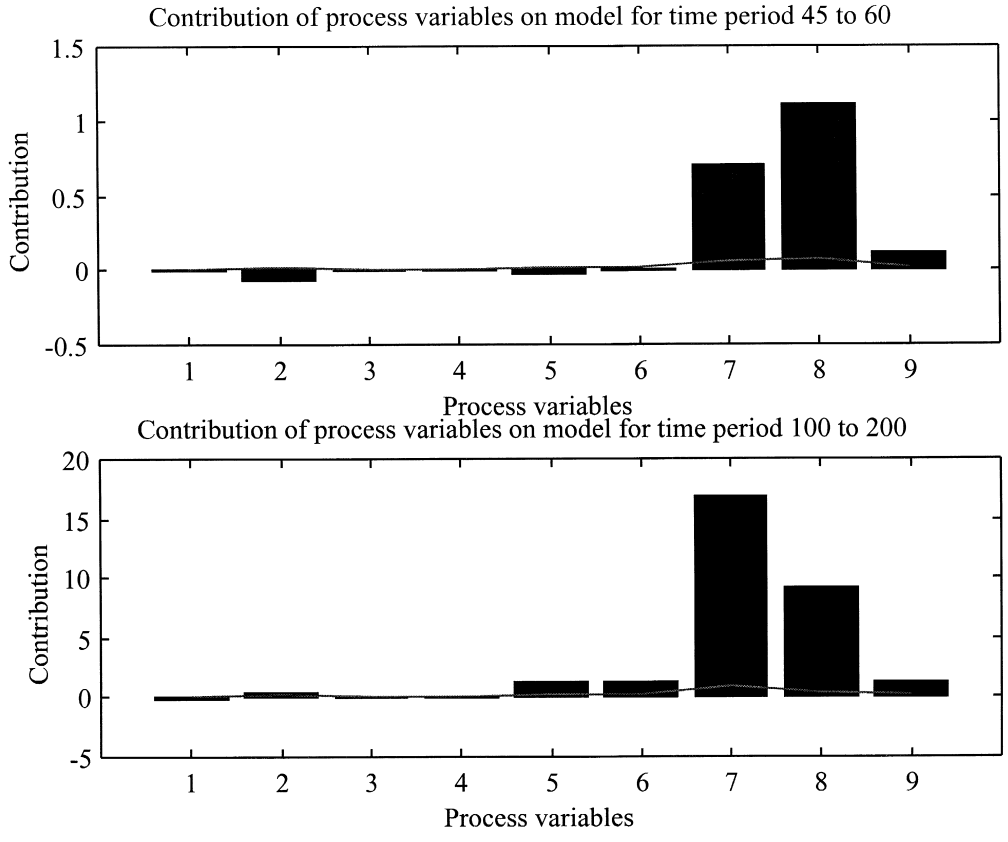

Fig. 12. Contribution of each process variable to *D*-statistic of batch 51 summed for time periods 45–60 (upper plot) and summed for time periods 100–200 (lower plot).

consecutive time points and $\mathbf{E}(I \times J)$ contains the residuals. The number of components R is usually much smaller than I and J. New measurements $x_{\text{new}}(J \times 1)$ can be projected on this model to see if the process is still in statistical control (see Eq. (2)). Then, $t_{\text{new}}(R \times 1)$ and $e_{\text{new}}(J \times 1)$ are the corresponding scores and residuals of the new batch run. Now the D- and Q-statistics can be calculated together with their corresponding limits as presented earlier in this paper.

If, in a continuous process, the D-statistic or the Q-statistic is above the confidence limit, contribution plots will directly show which of the J process variables was the main cause of the disturbance. It is not necessary to zoom in on a specific period of the process. This is comparable to the on-line monitoring case in a batch process. The contribution to the Q-statistic in a continuous process for process variable j is calculated as follows:

$$c_j^Q = \left(e_{\text{new},j}\right)^2 = \left(x_{\text{new},j} - \hat{x}_{\text{new},j}\right)^2 \tag{18}$$

where $e_{\text{new},j}$ is the residuals of a new measurement $x_{\text{new},j}$ and $\hat{x}_{\text{new},j}$ is the part of $x_{\text{new},j}$ explained by the process model. In this case, the smearing out of residuals over different process variables is also present. Therefore, the contributions to the Q-statistic should be interpreted with care.

In continuous processes, and also in on-line monitoring of a batch process, it is not necessary to square the residuals, because they are not summed. If the residuals are not squared, then they give information whether a process variable is too low or too high. If a rising trend is observed in the Q-statistic for con-

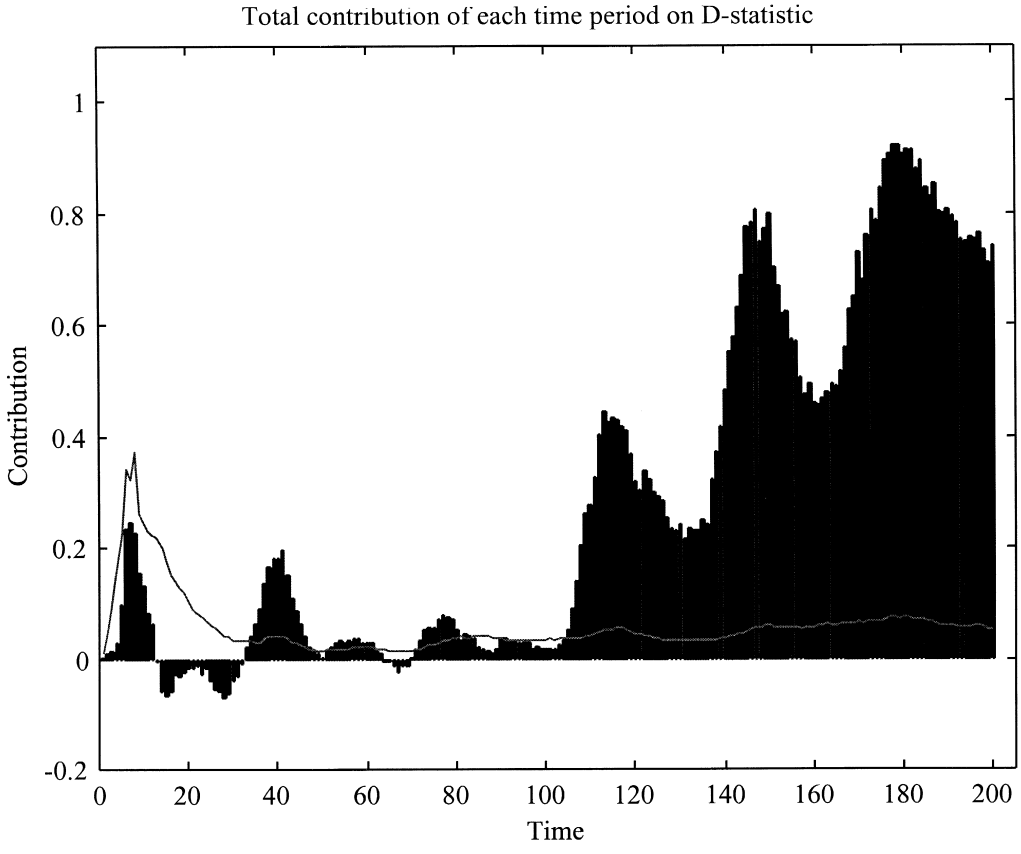

Fig. 13. Total contribution to D-statistic of batch 52 of all process variables summed for each time period. Control limits represent the contributions expected from NOC batches. This batch is known to have a disturbance coming in halfway the batch run.

secutive measurements, it is informative to study the trend for each process variable. MacGregor et al. [4] examined the difference in contribution between a measurement that is out of the confidence limits and the measurement at the start of the trend. Wise and Gallagher [28] studied the contribution of the process variables for each measurement during the trend. Applications of contribution plots to residuals for continuous processes can be found in Refs. [4,5,9,23, 24,28,29].

Contributions to the *D*-statistic for continuous processes can also be used in the same way as described for batch processes.

$$c_j^D = \mathbf{t}_{\text{new}}^{\mathrm{T}} \mathbf{S}^{-1} \left[x_{\text{new},j} \mathbf{p}_j \left(\mathbf{P}^{\mathrm{T}} \mathbf{P} \right)^{-1} \right]^{\mathrm{T}} \quad (19)$$

Control limits for the contributions to both *Q*- and *D*-statistic for continuous processes are calculated in the same way as presented earlier for batch processes.

3. Results

The use of the contribution plots presented in this paper is illustrated with a benchmark data set of a simulated semi-batch emulsion polymerization of styrene butadiene [17]. A semi-batch or fed batch process is a batch process that is not a closed system, e.g., a process where during the batch run, monomer is fed into the autoclave. Still the process has a specified duration. Meaningful disturbances such as impurities in the initial charge of the organic phase and in the butadiene feed to the reactor were added. Measurements were taken from flow rates, temperatures,

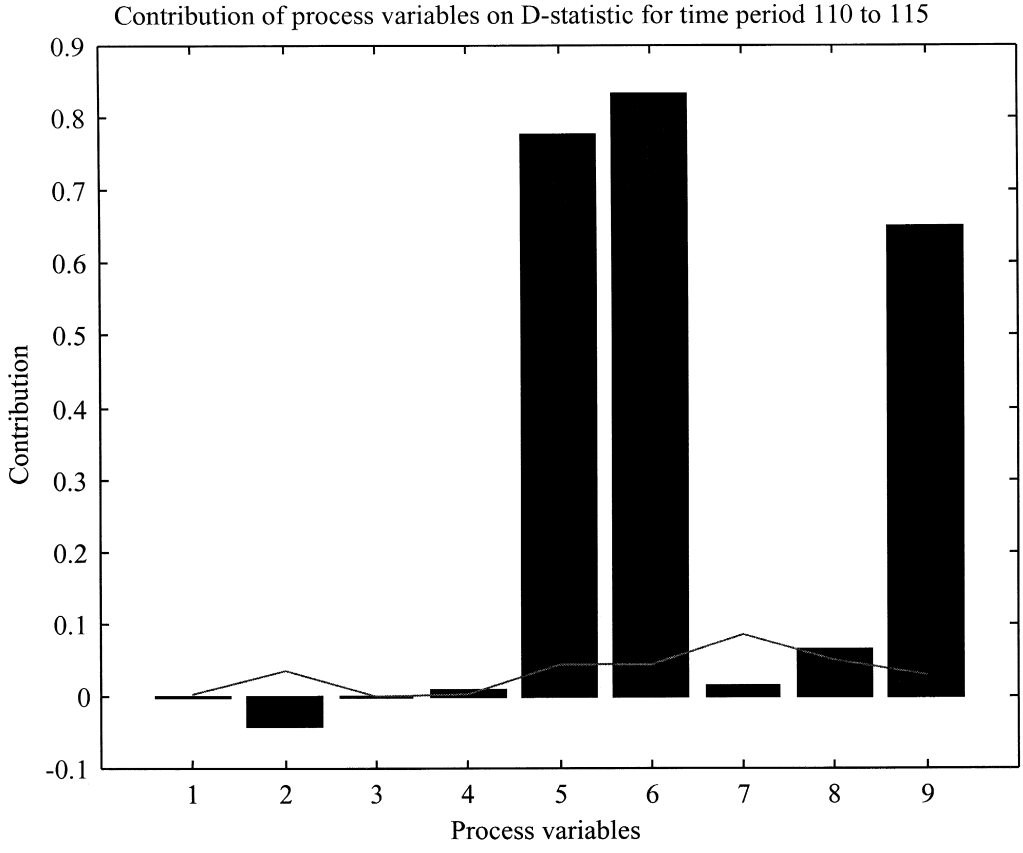

Fig. 14. Contribution of each process variable to *D*-statistic of batch 52 summed for time periods 110–115. Control limits represent the contributions expected from NOC batches.

density, estimates of the conversion and energy release. A detailed description can be found in literature [18]. Table 1 shows the process variables obtained within this data.

Fifty batches were simulated to construct the NOC data, by introducing typical variations. Three additional batches were simulated, one with normal conditions and two with product that was out of the specification range. One of the erroneous batches had an initial organic impurity contamination in the butadiene feed. The other erroneous batch had the same problem, but the contamination was higher and started halfway through the batch operation.

The NOC data were arranged in a three-way array $\underline{\mathbf{X}}(I \times J \times K)$ of $I = 50$ batches, $J = 9$ process variables and $K = 200$ time points. To describe the variation of the process variables around their average trajectories, each column of $\boldsymbol{X}$ was scaled to mean zero. Furthermore, each process variable was scaled to unit sum of squares. This type of scaling is called slab scaling [30,31].

The NOC data were modeled with a PARAFAC model. Three components were found to best fit the data. Although the 50 batches were simulated to come from NOC, two batches (12, 16) were fitted rather badly by the PARAFAC model. These two batches were removed from the NOC set, and the final $\underline{\mathbf{X}}(48 \times 9 \times 200)$ was used to develop the MSPC model.

The PARAFAC model with three components describes 21% of the total variance in the NOC data. This amount is small; however, such low percentages are often seen in modeling batch process data. Two of the three components of the PARAFAC model are rather correlated ($r = -0.89$). The PARAFAC model is not the optimal model with respect to described variance for these data, but it is used here to show that the generalized contributions can deal with correlated scores. The PARAFAC model is used to de-

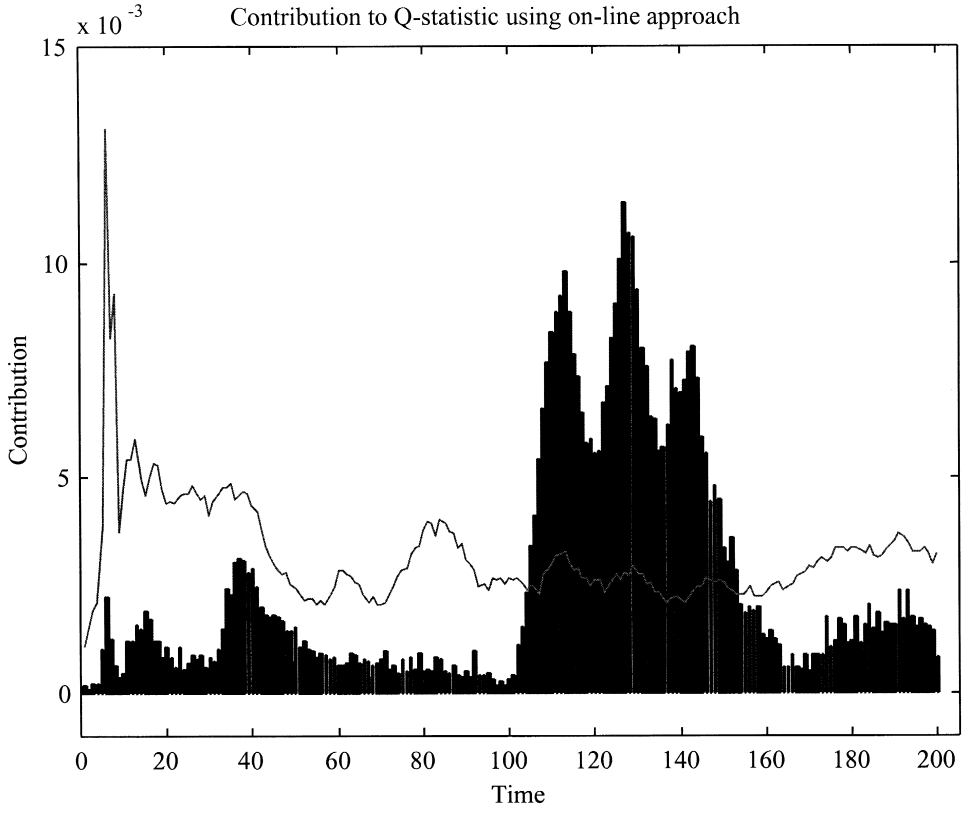

Fig. 15. Total contribution to Q-statistic of batch 52 of all process variables summed for each time period using the on-line approach. Control limits represent the contributions expected from NOC batches.

velop an off-line monitoring strategy using the *D*- and *Q*-statistic charts as described by Refs. [1,18]. Fig. 9 shows the *D*-statistic chart. The NOC batches (runs 1–50 without run 12 and 16) are all below the 95% confidence limit. These limits were determined using an *F*-distribution with R and $I-R$ i.e. 3 and 45 degrees of freedom. Batches 51 and 52, which are the erroneous batches are clearly detected as out of control. Batch 53, the extra batch obtained under normal conditions is clearly below the confidence limits. The *Q*-statistic chart in Fig. 10, shows again that all NOC batches are in control, and only batch 52, with the impurities entering the reactor halfway the process is above the confidence limit. The erroneous batch 51 is thus only detected in the *D*-statistic chart.

To find out what really happened in the erroneous batches, contribution plots can point out the specific process variables that show different behavior from the batches obtained under NOC. Fig. 11 shows the contribution of the process variables to the *D*-statistic for batch 51, the batch with the disturbance from the start. The contributions of all J process variables are summed for each time period. The control limits are calculated using a Jackknife procedure where for each NOC batch contributions were obtained. The control limits represent the mean plus three times the standard deviation of these contributions. In the beginning of the batch, the control limits are high. This is common in batch processes, where each run needs some time to stabilize. A lot of violations of the limits can be observed for this batch. After a small violation in the beginning of the run, the big problem starts at time 40. Then after the contributions come below the control limits, they go out for good after time period 90. It is better to zoom in on a specific

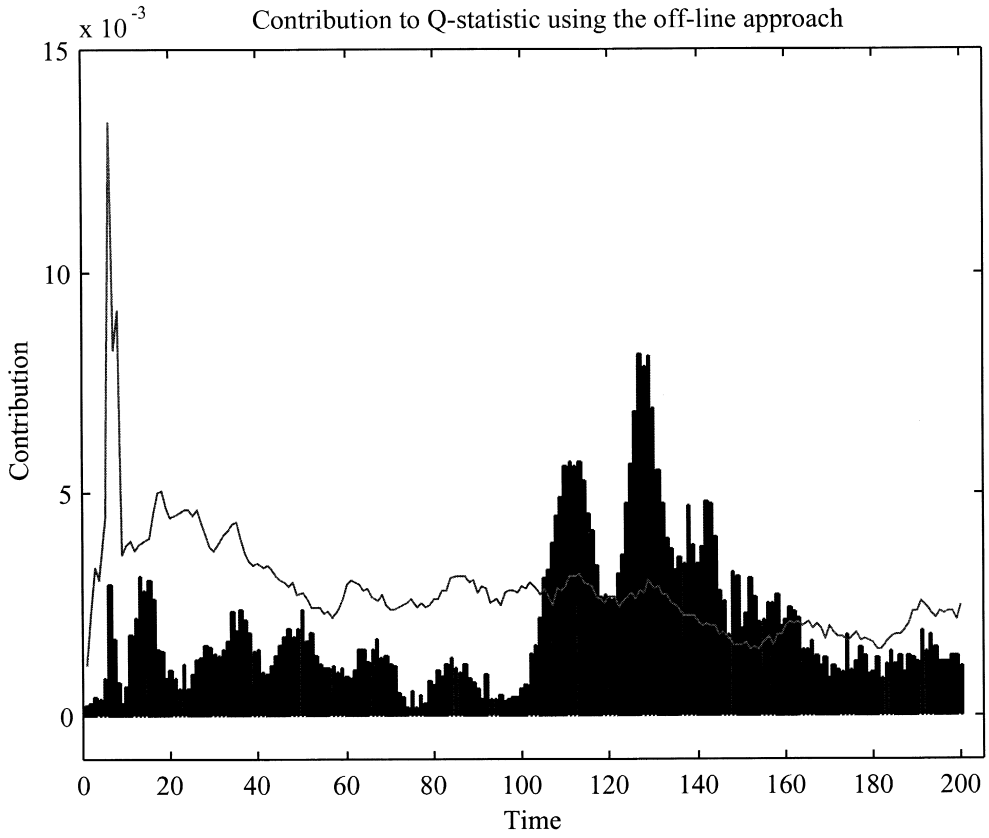

Fig. 16. Total contribution to *Q*-statistic of batch 52 of all process variables summed for each time period using the off-line approach. Control limits represent the contributions expected from NOC batches.

period where the batch started to go out of control. Fig. 12 shows the contributions when zoomed in on specific time periods during the run. At the top of this figure, the contributions of the process variables between time periods 45 and 60 are shown. This is the first period that is far above the control limits. Process variables 7 and 8 show the largest contributions. Process variable 9 is also somewhat higher than usual. Also for the last part of this specific run (period 100–200), process variables 7, (density of the latex), and 8 (conversion) are the ones that have the highest contributions. This is shown at the bottom of Fig. 12. The trajectories of these two process variables were lower than usual. Because variables 3–6 are behaving as usual, the cooling system seems to work properly. A possible explanation is that an impurity of the feed caused the decrease in conversion, which is known to be the case for this specific batch. The disturbance, which started at the beginning of the run, is not detected earlier, because the beginning of the run shows a large variation. This can be observed by the high control limits calculated for the contributions.

The second erroneous batch (batch run 52), was outside confidence limits both in the *D*-statistic and in the *Q*-statistic. Fig. 13 shows the contribution to the *D*-statistic of the process variables summed for each time period for this batch. After a small violation of the control limit in the beginning, the batch is really out of control after time period 100. It is known that at this time period, impurities were fed into the reactor. Zooming in on the period of the large violation (between time period 110 and 115), process variables 5, 6 and 9 have very high contributions (see Fig. 14). Fig. 15 shows the contribution of the process variables to the *Q*-statistic for batch 52. These contributions were obtained from residuals calculated according to the on-line approach. For a comparison, Fig. 16 shows the contributions determined from residuals obtained in the off-line mode. The scale and profile of the contributions is similar in both

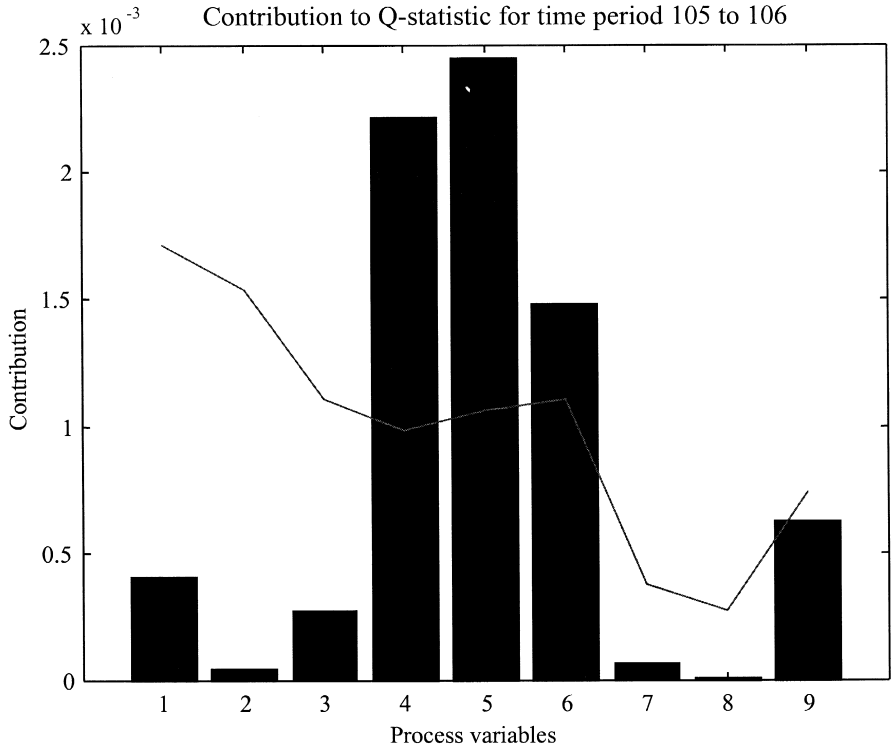

Fig. 17. Contribution of each process variable to *Q*-statistic for batch 52 summed for time periods 105–106. Control limits represent the contributions expected from NOC batches.

cases, but the on-line approach shows more clearly the time period where the process went out of control. Zooming in on the period where the *Q*-statistic was first outside the control limits (periods 105–106, Fig. 17), besides variables 5 and 6 that were already detected from the *D*-statistic contributions, variable 4, the reactor temperature, has a high contribution. When the trajectories of the process variables with high contributions (4, 5, 6 and 9) are examined, the reactor temperature is somewhat lower than usual for a small period. Directly after the drop in reactor temperature, the temperature in the cooling system (variables 5, 6) is higher than usual. This implies that the cooling system is compensating a temperature drop in the reactor. The instantaneous rate of energy release (variable 9) is also lower than normal, but no drop in feed temperature was detected. Hence, something in the reactor must have caused the temperature to go down. Again an explanation for this behavior is that impurities in the feed cause the reaction to slow down, which was the case in this specific batch.

Summarizing, the contribution plots presented in this paper are able to point to the specific period in the batch where the problem occurred. Zooming in on this period usually signals the process variables that are different from normal operating behavior. Control limits are used to show the relative contribution as compared to the contributions obtained from batches that were obtained under NOC.

4. Conclusions

In this paper, contributions to *D*-statistic and *Q*-statistic in statistical process monitoring are presented that can be used for any latent variable component or regression model to detect the specific process variable at a specific period during the run that caused the statistic to be out of control. The contributions to the *D*-statistic are not limited to models with orthogonal constraints. Furthermore, the issue of negative contributions to the *D*-statistic is discussed. For the contributions to the *Q*-statistic, a problem of smearing out of the residuals over time and over different process variables is addressed. The smearing out over time can be solved by assuming that the residuals were obtained in an on-line mode. The smearing out over different variables cannot be solved in this way, and therefore the contributions to the *Q*-statistic should be interpreted with care. Control limits are calculated for these contributions to show the relative contribution as compared to contributions of batches obtained under normal operating conditions. These limits help in detecting process variables that are really different from NOC behavior.

Acknowledgements

These investigations were supported by the Council for Chemical Sciences of the Netherlands Organization for Scientific Research (NWO-CW) with financial aid from the Netherlands Technology Foundation (STW). The authors would like to thank John MacGregor and Dora Kourti of McMaster University, Hamilton, ON, Canada for a fruitful discussion on this subject and for providing the simulated batch data.

References

[1] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (41) (1995) 41–59.
[2] J. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, Can. J. Chem. Eng. 69 (1991) 35–47.
[3] P. Miller, R. Swanson, C. Heckler, Contribution plots: a missing link in multivariate quality control, Applied Math and Computer Science 8 (1998) 775–792.
[4] J.F. MacGregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods, AIChE J. 40 (1994) 826–838.
[5] T. Kourti, J. Lee, J.F. MacGregor, Experiences with industrial applications of projection methods for multivariate statistical process control, Comput. Chem. Eng. 20: (1996) 745–S750.
[6] D. Neogi, C.E. Schlags, Multivariate statistical analysis of an emulsion batch process, Ind. Eng. Chem. Res. 37 (1998) 3971–3979.
[7] R. Boqué, A.K. Smilde, Monitoring and diagnosing batch processes with multiway covariates regression, AIChE J. 45 (1999) 1504–1520.
[8] K.S. Dahl, M.J. Piovoso, K.A. Kosanovich, Translating third-order data analysis methods to chemical batch processes, Chemom. Intell. Lab. Syst. 46 (1999) 161–180.
[9] C. Wikstrom, C. Albano, L. Eriksson, H. Friden, E. Johansson, A. Nordahl, S. Rannar, M. Sandberg, N. Kettaneh-Wold, S. Wold, Multivariate process and quality monitoring applied

to an electrolysis process: Part 1. Process supervision with multivariate control charts, Chemom. Intell. Lab. Syst. 42 (1998) 221–231.
[10] S. Wold, N. Kettaneh, H. Friden, A. Holmberg, Modelling and diagnosis of batch processes and analogous kinetic experiments, Chemom. Intell. Lab. Syst. 44 (1998) 331–440.
[11] E.B. Martin, A.J. Morris, M.C. Papazoglou, C. Kiparissides, Batch process monitoring for consistent production, Comput. Chem. Eng. 20 (1996) S599–S604.
[12] B.M. Wise, N.B. Gallagher, S. Watts Butler, D.D. White Jr., G.G. Barna, A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, J. Chemom. 13 (1999) 379–396.
[13] NAmICS (North American chapter of the International Chemometrics Society) Newsletter No. 19 (April, 1999).
[14] S.J. Wierda, Multivariate statistical process control — recent results and directions for future research, Statistica Neerlandica 48 (1994) 147–168.
[15] P. Nomikos, Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis, ISA Trans. 35 (1996) 259–266.
[16] T. Kourti, P. Nomikos, J.F. MacGregor, Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS, J. Process Control 5 (1995) 277–284.
[17] T.O. Broadhead, A.E. Hamielec, J.F. MacGregor, Dynamic modeling of the batch, semi-batch and continuous production of styrene–butadiene copolymers by emulsion polymerization, Makromol. Chem., Suppl. 10 (1985) 105–128.
[18] P. Nomikos, J.F. MacGregor, Monitoring of batch processes using multi-way principal component analysis, AIChE J. 40 (1994) 1361–1375.
[19] H.A.L. Kiers, Towards a standardized notation and terminoly in multi-way analysis, J. Chemom., in press.
[20] C.R. Rao, S.K. Mitra, Generalized Inverse of Matrices and its Applications, Wiley, New York, 1971.
[21] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate control charts for individual observations, J. Quality Technol. 24 (1992) 88–95.
[22] J.E. Jackson, G.S. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (1979) 341–349.
[23] T. Kourti, J.F. MacGregor, Multivariate SPC methods for process and product monitoring, J. Quality Technol. 28 (1996) 409–428.
[24] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemom. Intell. Lab. Syst. 28 (1995) 3–21.
[25] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[26] J.A. Westerhuis, P.M.J. Coenegracht, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, J. Chemom. 11 (1997) 379–392.
[27] D.J. Louwerse, A.K. Smilde, Multivariate statistical process control of batch processes based on three-way models, Chem. Eng. Sci. 55 (2000) 1225–1235.
[28] B.M. Wise, N.B. Gallagher, The process chemometrics approach to process monitoring and fault detection, J. Process Control 9 (1996) 329–348.
[29] J.F. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Eng. Practice 3 (1995) 403–414.
[30] R.A. Harshman, M.E. Lundy, Data preprocessing and the extended PARAFAC model, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. MacDonald (Eds.), Research Methods for Multimode Data Analysis, Preager, New York, 1984, pp. 216–284.
[31] R. Bro, A.K. Smilde, Centering and scaling in component analysis, submitted for publication.