ELSEVIER

Chemometrics and Intelligent Laboratory Systems 46 (1999) 197–206

Chemometrics and intelligent laboratory systems

PLS discriminant analysis with contribution plots to determine differences between parallel batch reactors in the process industry

D.J. Louwerse [a], Adriaan A. Tates [a], Age K. Smilde [a,*], Gerard L.M. Koot [b], H. Berndt [c]

[a] *Department of Chemical Engineering, Process Analysis and Chemometrics, Nieuwe Achtergracht 166, 1018 WV Amsterdam, Netherlands*
[b] *Shell Research and Technology Centre Amsterdam, Department of Engineering and Operational Support, Badhuisweg 3, 1031 CM Amsterdam, Netherlands*
[c] *Shell Nederland Chemie, PVC Manufacturing, Vondelingenweg 601, 3196 KK Rotterdam, Netherlands*

Received 21 April 1998; revised 4 June 1998; accepted 29 June 1998

Abstract

PLS discriminant analysis, applied to a PVC polymerisation batch process, is used to determine performance differences of parallel batch reactors. Weight contribution plots of time points and of variables are used to physically interpret the modelled differences; the main time points in which deviations occur and variables that cause the observed differences are assigned. A simple step-wise procedure is suggested to implement this method in the process industry. It was found that a systematic difference between the polymerisation time of the PVC batch reactors was caused by sensor failure or due to drifting thermocouples.

Keywords: PLS; Discriminant analysis; Contribution plots; Batch reactors; Sensor failure

Contents

* Corresponding author. Fax: +31-20-5256638; E-mail: asmilde@anal.chem.uva.nl

0169-7439/99/$ - see front matter
PII: S0169-7439(98)00185-3

1. Introduction

Batch production processes play an important role in chemical industry. Pharmaceuticals, biochemicals and a large number of polymers, for example, are often produced batch wise. The size of a batch reactor is restricted for a number of reasons like safety, heating and cooling capacity, etc. The capacity of a production site, therefore, often is enlarged by using batch reactors in parallel. When these parallel reactors are equally designed, the performance should also be comparable. When equal quality feed stock is used for these reactors, and when the process conditions are equal, the behaviour of the batch runs should be similar and the quality of the product should be comparable. In practice, however, the performance is often different, and the causes for these differences are not always obvious. In this paper, an example is presented with equal feed stock, similar reactors with a comparable history, but with an initially not understood difference in batch performance. It is important to assign the causes of these differences as they obviously come from deviating process conditions. These deviating process conditions can also produce products with different qualities.

Causes that give rise to systematic differences between parallel reactors with equal feed stock should be identified and if possible, eliminated. In this way, a more stable process with less variability in both the process conditions and in the final product quality can be obtained. If such a problem is at hand, it can be tried to locate the problem by comparing measured process variables univariately. However, it is difficult to observe differences in the behaviour of a variable between parallel batch reactors, especially if the variation from batch to batch is large. Common control charts, that are often used for continuous processes, cannot be used in this case. These charts only have one target value, where process variables of batch processes change as a function of time. For batch processes, the time trajectories of the variables have to be compared. This requires more advanced control charts [1,2]. As variables can influence each other, often a difference in process conditions will affect several variables simultaneously. When, for example, the temperature rises in a sealed vessel filled with an inert gas, the pressure will also rise. The pressure and the temperature can be monitored independently but the nature of the problem is bivariate, or in general multivariate.

Presently, large amounts of process data are collected and stored by powerful plant instrumentation and computers, but only a small part of this data is actually used. The information in this data can be utilised with a multivariate method to detect differences in process conditions more efficiently. The presented multivariate method takes into account all relevant process variables, as well as the time trajectory of these variables.

Detecting multivariate differences, however, is not sufficient. The process conditions that are causing these differences have to be determined. This means that not only the responsible (combination of) variables have to be assigned, but also at what time during the batch run they appear. Such a method then can be a useful tool in the process industry for process engineers to interpret the results. Unfold partial least squares (U-PLS) discriminant analysis is a candidate for that as it can easily model differences between batches. The modelled differences can be converted in weight contribution plots of variables and time periods in a similar way as contribution plots for PCA models [3,4]. This makes the interpretation of the model straightforward and easy to use.

2. Theory

2.1. Discriminant analysis with Unfold-PLS

Normal characters, including upper case characters, are scalars and refer to single data elements, or

can have a special meaning as will be explained in the text. Bold lower case characters refer to vectors. Bold upper case characters refer to two-way arrays or matrices. Bold underlined upper case characters refer to three-way arrays. The superscript ^{T}, attached to an array, like $\boldsymbol{X}^{T}$, refers to the transpose.

The main interest is to determine differences in process variations between parallel batch reactors. Batch process data, $\underline{\boldsymbol{X}}$, typically can be arranged in a three-way array of size $I \times J \times K$ (Fig. 1). For a number of I batches (first mode), J variables (second mode) are measured as a function of K time points (third mode). Multivariate methods like principal component analysis (PCA) and PLS can only deal with two-way arrays. For this purpose the three-way array $\underline{\boldsymbol{X}}(I \times J \times K)$ has to be converted into a two-way array $\boldsymbol{X}$. There are three possible ways to do this. Since it is the aim to compare batches from different reactors, batches can be regarded as objects and the conversion from three-way to two-way is made such that the batches remain objects in a separate mode. So the first mode of $\boldsymbol{X}(I \times M)$ is the batch mode, and the second mode is a combined mode for time and variables. A row i $(i = 1, \ldots, I)$ then represents a batch with elements, m $(m = 1, \ldots, M;\ M = J \times K)$ for all possible combinations of time and variable. A similar approach is used in batch MSPC [1,2]. Alternatively, the three-way array $\underline{\boldsymbol{X}}$ also could be analysed with three-way methods like N-PLS and PARAFAC [5,6].

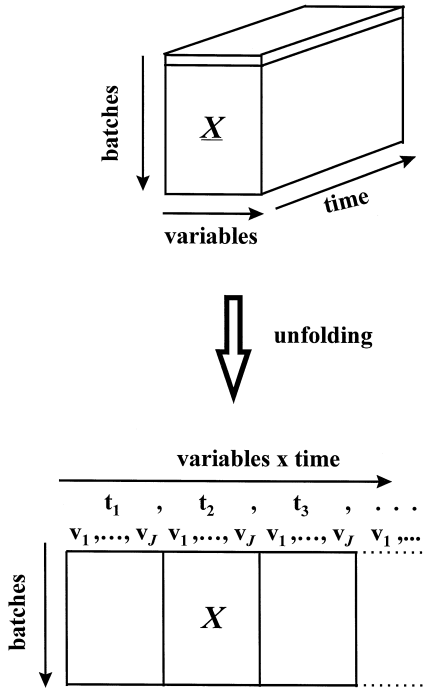

Fig. 1. The three-way array $\underline{\boldsymbol{X}}$ unfolded in a two-way array $\boldsymbol{X}$, where v_1, …, v_J denotes variable 1 until variable J, and t_1, t_2, t_3, … denotes the first, second, third, … time point.

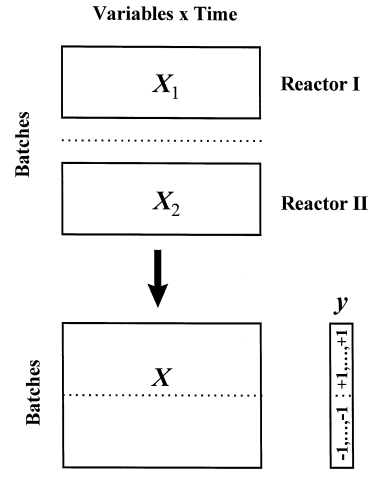

Fig. 2. Data arrangement to determine differences between the operation of batch reactors.

Wold et al. [7] described a PLS method to discriminate between groups of data, called PLS discriminant analysis (PLS-DA). Data from two batch reactors $\underline{\boldsymbol{X}}_1(I_1 \times J \times K)$ and $\underline{\boldsymbol{X}}_2(I_2 \times J \times K)$with I_1 and I_2 batches, J variables and K time points are combined to form $\underline{\boldsymbol{X}}(I \times J \times K)$. $\underline{\boldsymbol{X}}$ is unfolded into $\boldsymbol{X}(I \times M)$. For this combined matrix $\boldsymbol{X}$ a vector $\boldsymbol{y}$ of length I can be constructed, with elements being plus one or minus one, indicating the reactor origin. Fig. 2 schematically shows the arrangement of the data.

2.2. Data pre-treatment

The polymerisation time varies from batch to batch. This means that the data for different batches have a different number of time points K. Multivariate analysis requires the data to be stacked in a matrix, therefore K has to be set to a fixed number of time points. When the difference in time points between batches is not large the K time points can be calculated by linear interpolation. Aligning the batch

data with other methods, like dynamic time warping [8] also can be considered. In the presented case, the relative variation in polymerisation time is small. For this reason and also because it is a simple tool that can be used with a limited amount of pre-knowledge, linear interpolation is preferred here. First an arbitrary K is chosen, e.g., near the average number of time points of all batches. Suppose the data of batch i consists of $\tilde{k}_i = 1, \ldots, \tilde{K}_i$ time points. Then the increment of $\tilde{k}_i = 1, \ldots, \tilde{K}_i$ is stretched or shrunken to fit $\tilde{K}_i$ on K. Finally the values of all variables at time points $k = 1, \ldots, K$ are calculated by linear interpolating between the adjacent points of $\tilde{k}_i$.

A check should be made if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ contain abnormal behaving batches. Outlying batches have to be detected as they can influence the model more than normal operating batches. This check can be performed easily by modelling $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ each, with a multivariate technique like PCA. The size of the PCA model can be determined by cross-validation [9–11]. Hotelling statistics and the squared residuals can be used to detect batches that behave significantly different from the normal operating conditions (NOC) [2,12].

To equally weigh the variables, and to eliminate the non-linear behaviour of the time trajectory, every column of $\boldsymbol{X}$ is mean centred and scaled to unit variance. This approach also is used for monitoring batch processes with multivariate SPC charts [1].

2.3. PLS-DA weight contribution plots

The data shown in Fig. 2 can be modelled with PLS.

$$\begin{aligned} \boldsymbol{T} &= \boldsymbol{X}\boldsymbol{W}\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W}\right)^{-1} \\ X &= \sum_{r=1}^{R} \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} + \boldsymbol{E} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \\ y &= \sum_{r=1}^{R} \boldsymbol{t}_r q_r + f = \boldsymbol{T}\boldsymbol{q} + \boldsymbol{f} \end{aligned} \tag{1}$$

Where $\boldsymbol{T}$ is the score matrix, $\boldsymbol{P}$ the loading matrix for $\boldsymbol{X}$, $\boldsymbol{q}$ the loading vector for $\boldsymbol{y}$, $\boldsymbol{W}$ the weight matrix, $\boldsymbol{E}$ the residual matrix of $\boldsymbol{X}$ and $\boldsymbol{f}$ the residual vector for $\boldsymbol{y}$. The size of the PLS model, or the number of R latent vectors (LV), can be determined by cross-validation [13] and by taking into account the amount of explained variance in $\boldsymbol{X}$ and $\boldsymbol{y}$. Every process variable at every time point has a contribution to each LV, captured in the weight matrix $\boldsymbol{W}$ $(M \times R)$. When $\underline{\boldsymbol{X}}$ is unfolded according to Fig. 1, the first J elements in column vectors of $\boldsymbol{W}$ correspond to the J variables of the first time point; the elements $J+1$ to $2J$ correspond to the second time point, etc. Plotting all weight contributions, i.e., all elements of $\boldsymbol{W}$, is not very informative as the effect of time and variables are mixed. Plotting the weight contribution of one variable, summed over time is more informative. These weight contributions, a_{jr}, can be calculated for every LV according to

$$a_{jr} = \sum_{k=1}^{K} \left(w_{[(k-1)J+j]r}\right)^2 \tag{2}$$

This results in a matrix $\boldsymbol{A}$ $(J \times R)$ with elements a_{jr}, describing the weight contributions for each variable j and for each LV r, summed over all time points. In case of 15 variables, the weight contribution of the third variable to the first LV is: $a_{3,1} = w_{3,1}^2 + w_{18,1}^2 + w_{33,1}^2 + \ldots$. The same overall weight

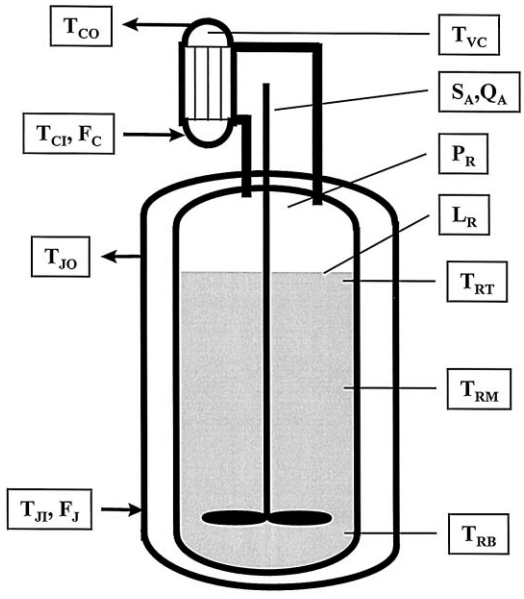

Fig. 3. The PVC batch reactor. The abbreviated process variables are listed in Table 1.

Table 1
Relevant process variables of a PVC batch process

Variable number	Variable name	Description
1	T_{CI}	Inlet temperature cooling water of the condenser
2	T_{CO}	Outlet temperature cooling water of the condenser
3	T_{VC}	VC gas temperature at the top of the condenser
4	Q_C	Calculated duty of the condenser
5	F_C	Amount of cooling water through the condenser
6	S_A	Agitator speed
7	Q_A	Power supply to the agitator
8	T_{RB}	Temperature of the reactor at the bottom
9	T_{RT}	Temperature of the reactor at the top
10	T_{RM}	Temperature of the reactor in the middle
11	T_{JO}	Outlet temperature of the cooling water through the reactor jacket
12	T_{JI}	Inlet temperature of the cooling water through the reactor jacket
13	L_R	Level of the batch reactor
14	F_J	Amount of cooling water through the jacket
15	P_R	Pressure of the reactor

factors can be calculated for every time point, summed over the variables.

$$b_{kr} = \sum_{j=1}^{J} \left(w_{[(k-1)J+j]r} \right)^2. \quad (3)$$

Likewise, this results in $\boldsymbol{B}$ $(K \times R)$ with elements b_{kr}, describing the weight contributions for each time point k and for each LV r, summed over all variables. In the same case, the weight contribution of the fourth time point to the second LV is: $b_{4,2} = w_{46,2}^2 + w_{47,2}^2 + \ldots + w_{60,2}^2$. Plotted columns of $\boldsymbol{A}$ indicate the weight contribution of individual variables to the differences in reactor performance for each separate LV. Likewise, the plotted columns of $\boldsymbol{B}$ show the time points responsible for the differences in reactor performance. More detail can be achieved by only considering a restricted part of $\boldsymbol{W}$, for instance, monitoring a specific variable close to a particular scheduling instance of the batch.

3. Experimental

3.1. PVC batch polymerisation process

Polyvinylchloride (PVC) is produced on a large scale by Shell on their production location in Pernis, located in The Netherlands. The vinylchloride monomer (VC) is polymerised in a water suspension. During the process, up to three phases exist: a water phase, a liquid VC phase, and a solid PVC phase. At the start of the batch, water, VC, suspension stabilisers and initiator is added to the reactor. The contents are stirred vigorously, so that a suspension of VC droplets in water is obtained. The reactor contents are heated to the polymerisation temperature. The heating is continued until the polymerisation reaction generates sufficient heat by itself. PVC is insoluble in water and only weakly soluble in VC, so it will precipitate quickly, forming a solid PVC phase inside the

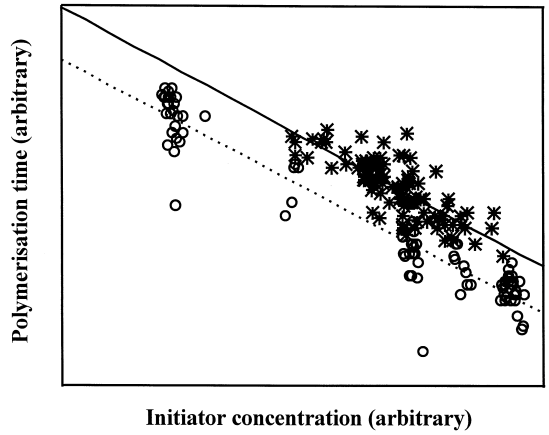

Fig. 4. Polymerisation time as a function of the initiator concentration, (∗) for reactor I fitted with a normal line; and (○) for reactor II fitted with a dashed line.

Action	Result
1. Construct a three-way array for both batch reactors by linear interpolating the variable responses for K time points	$\underline{\boldsymbol{X}}_1$ (101 x 15 x 60) $\underline{\boldsymbol{X}}_2$ (92 x 15 x 60)
2. Unfold $\underline{\boldsymbol{X}}_1$ and $\underline{\boldsymbol{X}}_2$	$\boldsymbol{X}_1$ (101 x 900) and $\boldsymbol{X}_2$ (92 x 900)
3. Check the data for outliers with PCA. This results in NOC data.	$\boldsymbol{X}_1$: 18 of 101 outlying batches $\boldsymbol{X}_2$: 9 of 92 outlying batches
4. Construct $\boldsymbol{X}$ with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and $\boldsymbol{y}$; mean centre and scale $\boldsymbol{X}$	$\boldsymbol{X}$ (166 x 900), $\boldsymbol{y}$ (166 x 1)
5. Model $\boldsymbol{X}$ and $\boldsymbol{y}$ with PLS, use cross validation and a table with variances in $\boldsymbol{X}$ and $\boldsymbol{y}$ explained by the model to determine R	$R = 1$
6. Calculate $\boldsymbol{A}$ and $\boldsymbol{B}$	$\boldsymbol{A}$ (15 x 1), $\boldsymbol{B}$ (60 x 1) Figure 7
7. Interpret results and fine-tune if necessary	Figure 7 and Figure 8

Fig. 5. Step-wise procedure of the PLS-DA and the result obtained in several stages of the method.

VC monomer droplets. The polymerisation takes place in the PVC phase and in the monomer phase.

A lot of heat has to be withdrawn from the process since the polymerisation reaction is highly exothermic. To assure the product quality, however, it is also important to keep the temperature on a constant target value. The excess of heat is withdrawn by a cooling jacket, surrounding the reactor, and by condensing monomer vapour to liquid in a condenser on top of the reactor. After a period of polymerisation, the monomer phase is no longer present and all remaining VC is present in the gas phase or in the polymer phase. The polymerisation continues and VC is absorbed from the gas phase, resulting in a decreasing pressure. The polymerisation is finally stopped by adding a killing agent.

The amount of initiator added to the batch is an important parameter. A high concentration results in a high reaction rate and hence in a short polymerisation time. However, a fast reaction also means that the heat produced per unit of time is large. For both safety and control reasons the used amount of initiator depends on the cooling capacity. Important process variables are: temperatures on several locations in the reactor, the condenser and the cooling system; flows of the cooling system; agitator speed and power supply; reactor level; and reactor pressure. Fig. 3 shows a schematic outline of the PVC batch process; the variables are listed in Table 1.

3.2. Differences in reactor performance

The polymerisation time is mainly determined by the initiator concentration and temperature. However, the same amount of initiator in different reactors sometimes results in a different polymerisation time. This is shown in Fig. 4, where the linear fit of the polymerisation time as a function of the initiator concentration is plotted for two comparable reactors. The process data of both reactors was collected for a

Table 2
Explained variance in $\boldsymbol{X}$ and $\boldsymbol{y}$ by the PLS model

	$\boldsymbol{X}$		$\boldsymbol{y}$	
LV	This LV	Total	This LV	Total
1	27.88	27.88	95.11	95.11
2	33.07	60.95	3.16	98.28
3	2.82	63.77	1.13	99.40
4	3.96	67.73	0.22	99.63
5	2.65	70.38	0.16	99.78

period of time, after having cleaned both reactors to eliminate a memory effect due to fouling. A PLS-DA model was calculated for the variables listed in Table 1, according to the action scheme of Fig. 5. Intermediate results are given in this figure and in Table 2. All data handling and calculations are performed in the program Matlab (The Math Works) with the PLS toolbox (Eigenvector Research).

4. Results

The data of the first reactor showed a group of eight and a group of four succeeding batches, and some individual batches that where behaving differently from the other batches. The data of the second reactor showed a group of four succeeding batches and some individual batches that behaved differently. It was coincidental that for both reactors an equal number of 83 batches were retained.

The scatter plot of the scores of the first two latent variables is shown in Fig. 6. It clearly shows two clusters, discriminated mainly by the first LV. The centre of both clusters is marked. PLS leave-one-out cross-validation showed more than five significant LVs. Fig. 6 and Table 2, however, show that the first LV mainly discriminates between both reactors. It explains more than 95% of the variation in $\boldsymbol{y}$, hence there is hardly any variation left over in $\boldsymbol{y}$ to explain for the other LVs. The main differences in reactor performance are modelled by the first LV. This LV, therefore, will be used to determine the variables that are mainly responsible for these differences.

The weight contribution plot of the variables (Fig. 7a) shows that the reactor temperature at the bottom (variable 8) is the dominant factor. The weight contribution plot of the time points (Fig. 7b) shows that the differences are not specifically located in time. A univariate plot of variable 8 for a batch near the multivariate average, shows clearly an overall higher temperature in the first reactor (Fig. 8a). This temperature effect, however, does not show up at the top of the reactor; Fig. 7a shows a marginal difference for this variable which is confirmed by Fig. 8b. The reactor temperature in the middle (Fig. 7a, variable 10) also shows a distinct difference, but surprisingly, this difference is reversed if compared to variable 8 (Fig. 8c). A simple conclusion that the temperature in one of the reactors is higher cannot be drawn. It can be questioned, however, whether these different temperature profiles in the reactors truly exist, or that the

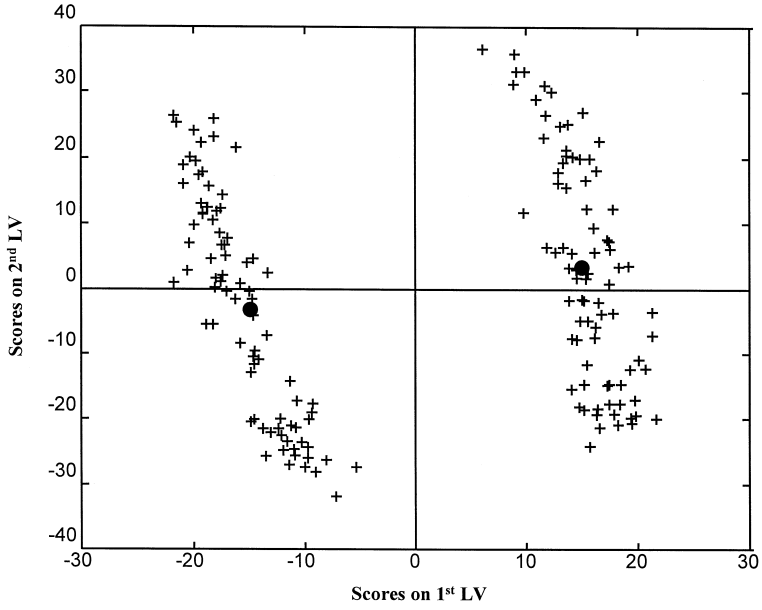

Fig. 6. Score plot of the first and second latent vector of $\boldsymbol{X}$. The average score of each reactor is denoted (●).

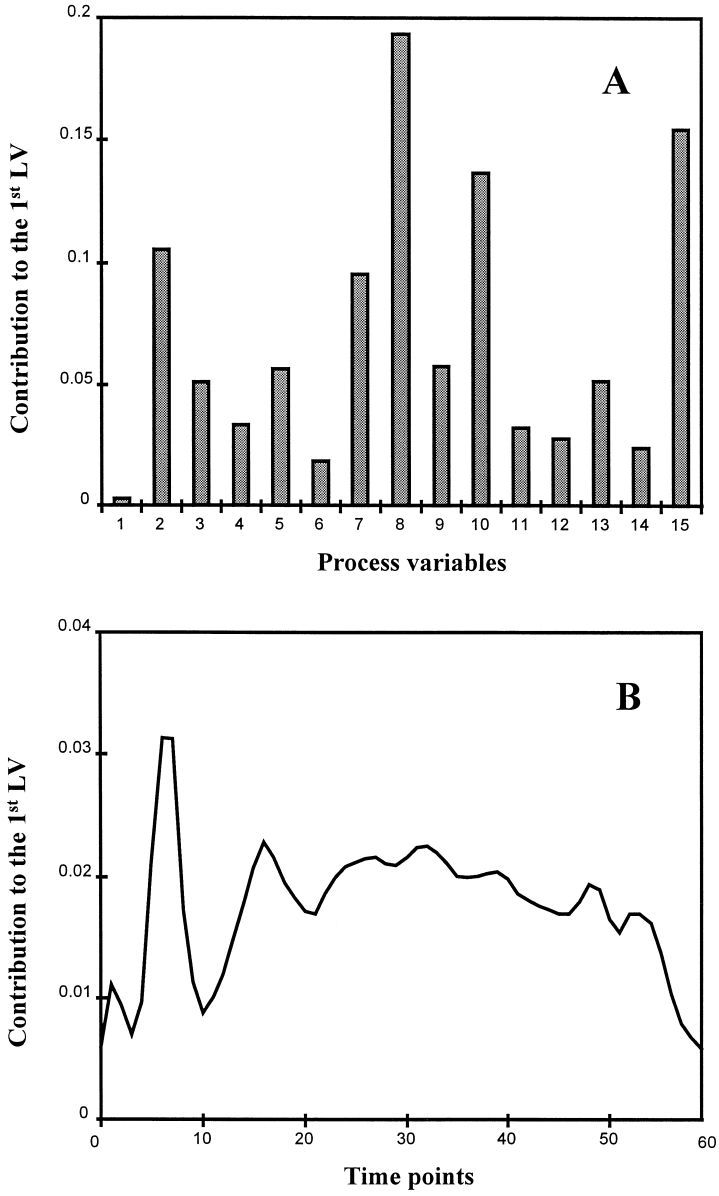

Fig. 7. Overall weight contribution to the first LV (A) assigned to process variables, and (B) assigned to time.

sensor readings are not accurate. The second dominant factor is the reactor pressure (variable 15). From the pressure profiles (Fig. 8d), it is clearly shown that the overall pressure in the second reactor is higher. In Fig. 4, it already was shown that for the same amount of initiator the polymerisation time in the second reactor is shorter. A higher overall temperature results in a higher overall pressure, a faster reaction rate and therefore, in a shorter polymerisation time. Given the existing differences in polymerisation times, the sensor readings of the pressure, the temperature at the top, and the temperature at the middle of the reactor, it can be concluded that the temperature of the second reactor is systematically higher. The only sensor

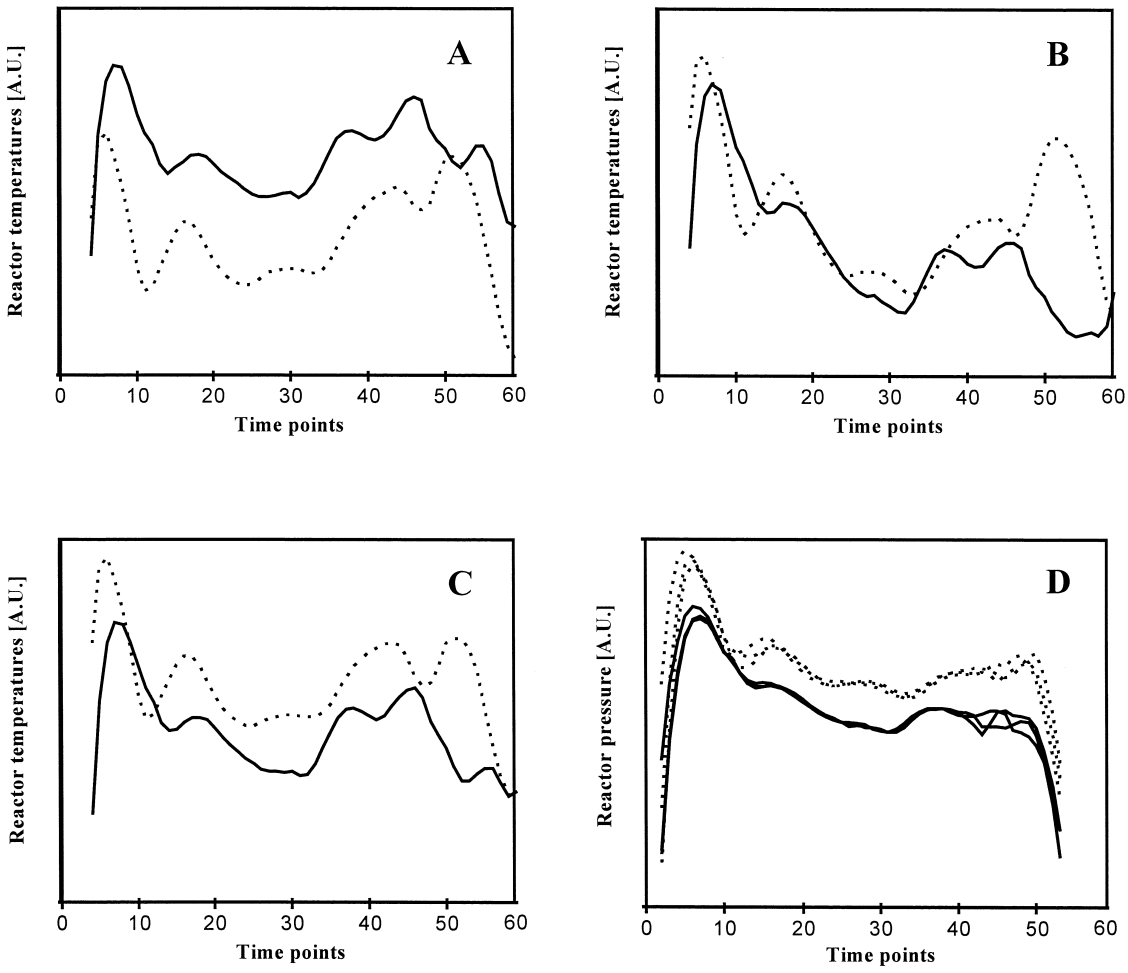

Fig. 8. Univariate plot of responses in time of (——) first and (- - -) second reactor. Temperature of the reactor at (A) the bottom, (B) the top and (C) the middle of a batch near the multivariate average. (D) The reactor pressure of three batches near the multivariate average of each reactor.

readings that deviate strongly from this, are those of the bottom reactor temperature. These deviating readings can only be explained due to faulty thermocouples. Next to the initiator influence, it is very likely that this is the main cause of the differences in polymerisation time between both reactors.

5. Discussion and conclusions

PLS-DA applied to batch processes proves to work well. A relatively simple stepwise procedure, like the one in Fig. 5, can be used to implement this method in the process industry. A big advantage can be obtained by using all relevant data simultaneously, instead of one variable at a time. The calculated weight contribution plots of the time points and of the variables make it possible for process engineers to physically interpret the PLS-DA modelled differences. The possibility to interpret the results in this way is essential for the acceptance of PLS-DA in the process industry. Dealing with all variables simultaneously and modelling them multivariably with PLS-DA, supplies additional information that cannot be attained by analysing variables univariately. Therefore,

PLS-DA is a supplementary tool which can be very valuable in the process industry.

It was found that, in case of the investigated PVC reactors, there was a systematic difference between the reactor polymerisation time due to a difference in the reactor temperatures. As the temperature targets of both reactors were equal, this can only be caused by sensor failure or due to drifting thermocouples. By using PLS-DA techniques, this could be established very quickly.

Acknowledgements

We thank Shell Nederland Chemie for providing the batch data of the PVC polymerisation process and for their support in this study.

References

[1] P. Nomikos, J.F. MacGregor, Monitoring batch processes using multiway principal component analysis, AIChE J. 40 (1994) 1361–1375.
[2] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.
[3] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemometr. Intell. Lab. Syst. 28 (1995) 3–21.
[4] P. Miller, R. Swanson, C. Heckler, Contribution plots: A missing link in multivariate quality control, to appear in Appl. Math. Comp. Sci. 8 (1998).
[5] R. Bro, H. Heimdal, Enzymatic browning of vegetables. Calibration and analysis of variance by multiway methods, Chemometr. Intell. Lab. Syst. 34 (1996) 85–102.
[6] A.K. Smilde, D.A. Doornbos, Three-way methods for the calibration of chromatographic systems: comparing parafac and three-way PLS, J. Chemometr. 5 (1991) 345–360.
[7] S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, in: B.R. Kowalski (Ed.), Chemometrics, Mathematics and Statistics in Chemistry, 1984, pp. 17–95.
[8] A. Kassidas, J.F. MacGregor, P.A. Taylor, Synchronization of batch trajectories using dynamic time warping, AIChE J. 44 (1998) 864–875.
[9] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (1978) 397–405.
[10] H.T. Eastment, W.J. Krzanowski, Cross-validatory choice of the number of components from a principal component analysis, Technometrics 24 (1982) 73–77.
[11] D.J. Louwerse, A.K. Smilde, H.A.L. Kiers, Comment on Eastment and Krzanowski, Technometrics (1982), submitted.
[12] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate control charts for individual observations, J. Quality Technol. 24 (1992) 88–95.
[13] M.J. Stone, R. Stat. Soc. B 36 (1973) 111–133.