J. Iran. Chem. Soc., Vol. 8, No. 1, March 2011, pp. 209-222.

JOURNAL OF THE
Iranian
Chemical Society

On-Line Monitoring of a Continuous Pharmaceutical Process Using Parallel Factor Analysis and Unfolding Multivariate Statistical Process Control Representation

M. Kompany-Zareh
Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran

(Received 24 January 2010, Accepted 2 July 2010)

Archive of SID

On-line high performance liquid chromatography (HPLC) was used to monitor steady state reactions in three reactors (K = 3) over 48.0 h. Different numbers of chromatograms, with J = 1981 retention time points, were recorded for each of the three reactors. Peaks for each chromatogram were baseline corrected and aligned using correlation optimized warping (COW). To make a complete three-way data set of I = 266 chromatograms a cubic Hermite interpolation was performed. The applied bilinear multivariate statistical process control (MSPC) method included the unfolding PCA and the trilinear technique was PARAFAC. Unfolding in reactor (K) mode was the most informative. D-charts and Q-charts were applied to the data to determine samples which were out of control. Confidence limits were then applied to the D and Q-charts and variables with different behaviours from that encapsulated within the reference data set were located. Both bilinear and trilinear methods were found to be useful for process analysis.

Keywords: Multiway on-line continuous process monitoring, Parallel factor analysis, Unfolding multivariate statistical process control, Fault detection, High performance liquid chromatography

INTRODUCTION

Continuous processes are of considerable importance for manufacturing of valuable products in many sectors of industry. These include pharmaceuticals, fine chemicals, polymers, crystallization and food. With increasing global competition it is of critical importance to ensure that the product of a continuous process is of consistent quality and at maximum yield. Thus, a highly desirable aspect of process operation is process performance monitoring. The aim of process monitoring is to achieve early warning of changes in process operation that may result in off-specification product properties and low quality production. It may be possible to take corrective action to recover the continuous process and prevent non-conforming product to be manufactured. Manufacturing must proceed in a reliable and robust manner with an acceptable false alarm rate.

Optical techniques, such as Near-Infrared, Mid-Infrared, Ultraviolet-Visible and Raman spectroscopy are common for on-line and/or off-line monitoring of chemical reactions and processes. In off-line process monitoring the data analysis is carried out after the completion of the process. These techniques are robust and easy to interface with fibre optic probes that could be inserted into reaction mixtures. Composition of the reaction mixtures can be monitored online and regularly with time, using a variety of chemometric methods for the resolution of the spectral profiles. Principal components analysis (PCA) [1-3], partial least squares (PLS) [2,3], kinetic modelling [4], iterative target transformation factor analysis (ITTFA) [5], mixed kinetic and multivariate

*Corresponding author. E-mail: kompanym@iasbs.ac.ir methods [6] and multivariate statistical process control [7-10] are among the chemometric methods employed.

Spectroscopic techniques are useful for measuring bulk changes in a system, but are not suitable for the detection of trace components, such as impurities, during the reaction. Impurities may often be chemically and structurally similar, and as a result, be spectroscopically similar to main reactants, especially if they are isomers. Usually, impurities can be detected by off-line methods such as liquid chromatography mass spectrometry (LCMS) and high performance liquid chromatography, sometimes a few days after the completion of the process.

High sensitivity, specificity and fast chromatographic procedures are the advantages of HPLC. The method can provide a proper monitoring complementary to the existing spectroscopic techniques. It is particularly powerful in impurity detection. On-line HPLC instrumentation is available commercially [11] and is finding applications in electronics [12], nuclear power [13], pharmaceutical [14] and chemical [15] industries including fermentations, semiconductor manufacturing and chemical synthesis. The present research makes use of on-line HPLC together with multivariate statistical process control to monitor reactions.

On-line HPLC generates a large amount of data. The results from efficient analysis of data reflect the state of process at the time of analysis. Before applying MSPC to the chromatographic data, several pretreatment steps have to be followed. Pretreatments include alignment and baseline correction. MSPC approaches can be applied to the pretreated data. The approach allows the data to be used for generation of an MSPC model to monitor and detect the changes in processes in real-time.

The data set in this work is obtained from a six-stage steady state process and involves three out of six stages over the period of 48 h, using rapid on-line HPLC. Multivariate statistical process control tools for monitoring of processes include bilinear techniques of multi-way principal component analysis (MPCA), multi-way partial least squares (MPLS) [7,16-18] and Hierarchical PCA [19] and, more recently, trilinear family of methodologies including direct trilinear decomposition (TLD) [20], parallel factor analysis (PARAFAC) [20-25] and Tucker3 [26]. Westerhuis *et al.* [27] compared PARAFAC, Tucker3 and MPCA on a simulation and an industrial polymerization process. MPCA includes an unfolding step in which the three way cube of data, *I*×*J*×*K*, is unfolded into a two-way *IJ*×*K* (or *I*×*JK*, or *J*×*KI)* data as illustrated in Fig. 1. PCA are then applied to the two-way data. In three-way approaches unfolding is not required and the cube of data is decomposed to three loading matrices (modes).

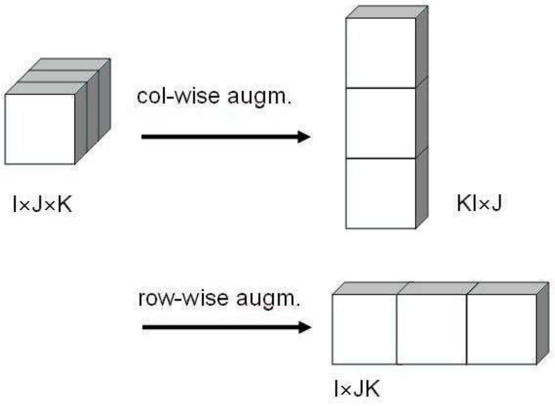

Fig. 1. Reactor(*K*)-unfolding-column-augmented and reactor (*K*)-unfolding-row-augmented data for MPCA.

In a number of research reports, the bilinear family of techniques were shown to be more appropriate for batch monitoring, since a greater percentage of variability in the data was explained with fewer latent variables. It might be argued, however, that the level of variability explained by a model is not the sole criterion for defining the most appropriate model for performance monitoring. Other issues to be considered include the time to detect a process change and the number of spurious alarms. In an application to semiconductor etch [20], it was shown that application of PARAFAC was slightly more sensitive to the detection of faults compared to MPCA, TLD and PCA. Based on the works and discussions in the two previously mentioned papers, and other similar works, PARAFAC and some bilinear techniques are investigated here for on-line process performance monitoring.

Westerhuis *et al*. [27] compared PARAFAC, Tubcker3, observation unfolding MPLS, and batch unfolding MPCA on two data sets and concluded that batch unfolding MPCA was the preferred method. Wise *et al*. [20] concluded that PARAFAC functions slightly better than unfolding MPCA for

Archive of SID

www.SID.ir

online fault detection of semiconductor production process. Louwerse and Smilde [28] and Smilde [23] outlined the theoretical aspects and comparative results of Tucker3, PARAFAC and batch unfolding MPCA and concluded that no clear conclusions could be drawn about the most effective method for process analysis.

Other multivariate analytical methods have also been proposed recently. They include moving window PCA [29], time varying state space modelling [30], multiway kernel PCA [31], independent component analysis [32], and stage based PCA [33]. It is clear that researchers have not come to an agreement as to which approach is most effective for batch analysis. The purpose of this study is to substantiate the appropriacy of both bilinear and trilinear based approaches to shed light on the controversial issue of monitoring a continuous process in the pharmaceutical industry.

The philosophy of on-line MSPC is similar to that of univariate SPC. A model is firstly developed based on the data collected during the manufacture of an acceptable product. This model then forms the basis for the on-line monitoring scheme, and the appropriate matrices are calculated including the action and warning limits. For a PARAFAC representation, the issues of robustness and reliability of an on-line monitoring scheme are investigated by confidence limits.

Traditionally, the location of variables (retention times) indicative of causing a change in the operational conditions has been investigated through contribution plots [34,35]. Application of these plots has also been reported for trilinear approaches [36]. In this study, both bilinear and trilinear approaches are tried on the data sets of an industrial continuous process. The proposed on-line monitoring methodology is investigated by its application to the data collected from three parts of a six-part reactor.

EXPERIMENTAL AND METHODS

Data Arrangement and Pre-processing

The reactions were monitored for 48.0 h. The sampling was irregular in time ranging from 5 to 50 min between each reaction sampling time. The outcome was 57, 177 and 186 samples from three reactors. Over the period of 2.5 min, with sampling interval of 0.0012 min, an HPLC chromatogram at 245 nm was recorded for each reaction sample. In this way, the number of data points from each chromatogram run was 2040. Three raw data matrices $\mathbf{X}_1$(51×2040), $\mathbf{X}_2$(177×2040) and $\mathbf{X}_3$(186×2040), whose rows and columns corresponded to reaction times and HPLC elution times, respectively, were the available data. It should be noted that sampling was uneven in the reaction time.

Data Pre-Processing

A number of steps were followed before the application of multivariate methods to process monitoring.

HPLC Baseline correction. Baseline correction was the first step. It involved the following stages:

(i) The background estimation. Here the moving windows of length *s* were defined for each chromatogram at sampling point *i* in reaction time, $\mathbf{x}_i$. Each window started *d* data points later than the previous window. So, the first window was from point 1 to *s*, the second from *d* to *d*+*s*, the third from 2*d* to 2*d*+*s* and so on. To ensure that there are regions of background within each window, value of *s* must be larger than the largest peak width detected. However, a too large window size results in less logical character of estimated background points. Sufficiently large vector of background points was obtained using small value of *d*. The value of background in the window centred on point *j*, b_{ij}, was estimated by selecting a 10% quantile value of the numbers in the window. For a chromatogram of *J* data points, $(J - s)/d + 1$ background points were obtained. In this study, values of $s = 50$ and $d = 10$ resulted in 200 background points for each chromatogram.

(ii) Interpolation of background points to the baseline. Elements of background vector for the chromatogram *i*, $\mathbf{b}_i$, were interpolated to provide an estimate of the baseline for each data point in the chromatogram, $\mathbf{f}_i$. A shape preserving piecewise cubic Hermite interpolation was utilized [37,38].

(iii) Subtracting the background. The *i*th background corrected chromatogram was obtained by subtraction of the background from the signal in the final step, $\mathbf{g}_i = \mathbf{x}_i - \mathbf{f}_i$.

Peak alignment. The next step was to align chromatograms in each of the three data matrices using correlation optimized warping (COW) [39,40]. This method involves breaking each chromatogram into segments, and then, warpping the data to obtain maximum correlation between each individual chromatogram and a target. The steps

Archive of SID

are as follows:

(i) The first chromatogram was defined as the target chromatogram for the second one. The aligned second chromatogram was the target for the third one. To validate the method, using a shuffle test, the sequence of chromatograms was changed. COW alignment was repeated for the changed data, and the same results were obtained. It showed that the choice of target chromatogram was not critical in this case.

(ii) The chromatogram to be aligned and the target were divided into similar number of segments. Dividing the number of points in the target chromatogram, *J*, by the length of each segment, J_r, the number of segments could be obtained. The length of each segment was allowed to be within the range of $J_\mathrm{r} + \delta$ (δ was the slack parameter). In this work, the values of J_r and δ were 100 and 50 point, respectively. If the number of data points in a segment of the chromatogram to be aligned was not equal to the number of data points in a specific segment of the target chromatogram, linear interpolaton of the sample was utilized to make the number of data points equal to the number of data points in the corresponding segment of the target.

(iii). Initially, the length of each segment in the chromatogram to be aligned was equal to J_r. In an iterative manner, by changing the lengths of the segments in the chromatogram to be aligned, the segments were shifted and rescaled to optimal lengths. So, correlation coefficient between each segment *m* was as close as possible to one. $\mathbf{r}_\mathrm{m}$ and $\mathbf{s}_\mathrm{m}$ were the corresponding segments in the target and chromatogram to be aligned, respectively [39,40].

Fixed Size Moving Window Evolving Factor Analysis (FSMWEFA)

One of the most valuable techniques to be used by chemometricians is principal component analysis (PCA) [1-3]. It decomposes the data into scores and loadings which are the principal factors. Local information in process monitoring may include the number of reactants changing within a time window, periods where faults occur in the process, impurities appearing at different times and the region where the process is in steady state. In the case of two-way data, obtained from unfolding a three-way data, fixed size moving window evolving factor analysis [3,41-46] extracts the local information through application of a series of PCA on submatrices from the original data set, in an evolutionary manner. The method gives better resolved information about local regions. *I-w* sets of eigenvalues were obtained, which when ordered according to sampling time, allowed us to explore how the process was changing with time. *w* was the window size and was selected as 10 in this work. FSMWEFA could only distinguish between the non-steady state and steady state normal operating condition (NOC) regions. The chromatograms from NOC regions, then, were employed for modeling using statistical control charts combined with other quality criteria.

Parallel Factor Analysis

The trilinear methodology of parallel factor analysis (PARAFAC) is based on the principle of proportional profiles [47]. It is one of the simplest three-way generalizations of the traditional multivariate statistical technique of factor analysis. The PARAFAC model was independently proposed by Carroll and Chang [48], who named the model CANDECOMP (canonical decomposition), and Harshman [49], who proposed the acronym PARAFAC. The PARAFAC model of a three-way array can be described as the sum of the outer products of a set of loading vectors [50]. Each set comprises three vectors according to the three modes of the original data. For example, in a continuous process the modes are time, variable and location (reactor number).

For the mathematical interpretation of a PARAFAC model, let $\mathbf{X}(I\times J\times K)$ denote a data matrix comprising *I* times, *J* variables and *K* time points. The PARAFAC decomposition is given by

$$x_{ijk} = \sum_{l=1}^{L} a_{il} b_{jl} c_{kl} + e_{ijk} \qquad (1)$$

where *L* is the number of factors included in the model, e_{ijk} is the error between the original data and the data projected down onto the model and a_{il} (i = 1, …, *I*) is the element of the *l*th factor for one of the three dimensions. Likewise, b_{jl} (j = 1, …, *J*) and c_{kl} (k = 1, …, *K*) define the elements of the factors for the other two dimensions. The objective of the model is to minimize the sum of squared errors:

$$\min\left(\sum_{i,j,k=1}^{I,J,K} e_{ijk}^2\right) \qquad (2)$$

Archive of SID

www.SID.ir

A schematic diagram of a PARAFAC model (Eq. 1) is shown in Fig. 2. Each block on the right side of Fig. 2 is related to the execution of an outer product between three loading vectors for one factor. *L* factors are considered in total. For a continuous process, **a** represents the mode of sampling time, **b** represents the mode of process variable (chromatographic retention time) and **c** the mode of location of measurement (reactor number).

MSPC Charts

Q and D charts. Determining whether a reaction is deviating from a set of conditions that are defined to be NOC conditions is the aim of MSPC charts [6-8]. In our system, a region of process including *n* chromatograms ($n < I$) was selected and defined as NOC. NOC region must meet the quality requirements. For instance, certain products must be more than a level and some impurities must be less than a level, in this region. The FSMWFA results, also, must show that they are within the steady state region. The control (reference) region $\mathbf{X_r}$ consisting of *N* rows, *J* columns and *K* layers, corresponding to the steady state, was decomposed by PARAFAC.

$$\mathbf{X_r}_{\,(I\times JK)} = \mathbf{A_r}\,(\mathbf{C_r} \otimes \mathbf{B_r}) + \mathbf{E} \qquad (3)$$

where (⊗) is the Kronecker product sign [24]. $\mathbf{A_r}$, $\mathbf{C_r}$, and $\mathbf{B_r}$ model the systematic variations. **E** describes the residuals not explained by the model. No scaling, mean centering, and/or standardization were applied on the data in this work. *L*, which is much less than *n* or *J* in most cases, is the number of significant components in the model. In the context of process control, no overall accepted criterion exists for choosing a proper value for *L*. However, a good approach is to select it so that most of the samples in the NOC region are within predefined control limits. In this study, the proper value of *L* was set to 3. On the other hand, the data are from three reactors that ideally, and when in the steady state, result in three significant factors. Based on this approach, the selection of three significant factors seems proper.

Two common statistical indicators in MSPC are D-statistic and Q-statistic [6-9]. D-statistic shows the distance of a part of process to the center of NOC samples. It determines whether a specific sample has a systematic deviation from the steady state region. When trying to fit the new samples to the model which is obtained from the NOC samples, Q-statistic can be used to estimate the residuals. The control limits for these statistics can be computed as described below, when both are assumed to follow the normal distribution. Values of relevant statistic in different sampling time points, $I = 1$ to *I*, produces the MSPC chart. 95% of samples that are under control are expected to fall within the 95% control limit, which is widespread. If a number of sequential samples are outside the limit, presence of a problem in the process under consideration is highly probable. Graphical indication of the limits, which is popular, allows the user to distinguish the problem more conventionally and rapidly. In this study, 99% (action) and 95% (warning) control limits are considered.

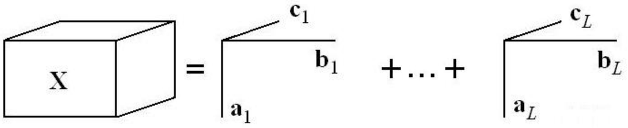

Fig. 2. Schematic form of the PARAFAC model.

The objective of the proposed D-statistic algorithm is to approximate a score vector at each sampling time point $i = 1,\ldots,I$, using the matrix of observations in each time point *i*. The first step is to calculate scores based on the reference (control) set of samples. The length of score vector at sampling point *i*, $\mathbf{t}_i$, is equal to the number of selected factors, *i.e.* *L*×1. For all the samples, reference and others (such as the newly measured samples) the score vector is then obtained by projecting the monitored observations from the new sample onto the reduced space. Hence, at sampling time point *i*, $\mathbf{t}_\mathrm{i}$ is given by

$$\mathbf{t}_\mathrm{i} = (\mathbf{P_r}^{\mathrm{T}} * \mathbf{P_r})^{-1} * \mathbf{P_r}^{\mathrm{T}} * \mathbf{x}_i = \mathbf{P_r}^{+} * \mathbf{x}_i \qquad (4)$$

where $\mathbf{P_r}$ is the loading matrix from reference samples, with *J*×*K* rows and *L* columns. Reference data in this study was obtained from sampling time point 69 to 95. $\mathbf{x}_\mathrm{i}$ represents the vectorized observations from i_{th} sample, with *J*×*K* rows and one column. $\mathbf{P_r}^{+}$ is pseudo-inverse of $\mathbf{P_r}$ and for a PARAFAC model, $\mathbf{P_r}$ is calculated as a Kronecker product , ⊗, as follows:

$$\mathbf{P_r}(:,l) = \mathbf{c}_{\mathbf{r}l} \otimes \mathbf{b}_{\mathbf{r}l}\ ,\ l=1,\ \ldots,\ L \qquad (5)$$

www.SID.ir

$\mathbf{b}_{\mathbf{r}l}$ and $\mathbf{c}_{\mathbf{r}l}$ are the loading vectors for factor l in the retention time (or variable J) and reactor number (K) modes, respectively.

In the next step, the process measurements at time point i, $\mathbf{x}_i$, can be estimated from score vector at time point i, $\mathbf{t}_i$. This is obtained by calculating the product of scores $\mathbf{t}_i$ and the loading matrix corresponding to the reference set:

$$\mathbf{x}_i = \mathbf{P_r} * \mathbf{t}_i \tag{6}$$

where $\mathbf{P_r}$ is calculated as Eq. (5). Finally, the score vector for any new sampling time can be estimated in a similar fashion, using the background corrected and aligned data from that time, and the $\mathbf{P_r}$ (JK×L) which is the loading matrix of the reference data set. The above procedure can be adopted at each sampling time point i of the process to obtain the corresponding score $\mathbf{t}_i$. The score vectors can then be combined to give a matrix which represents the time trajectory of the on-line scores. For each time point i there is an associated value for the Q-statistic and Hotelling's T^2. An estimate of Q-statistic at time point k can be made from the following equations:

$$\mathbf{err}_i = \mathbf{x}_i - \mathbf{P_r} * \mathbf{t}_i \tag{7}$$

$$Q_i = \sum_{s=1}^{JK} \mathbf{err}_i(s)^2 \tag{8}$$

The control limits for the Q-statistic are calculated from the χ^2 distribution and are given by:

$$Q\_\sim g\,\chi_{\mathrm{h},\alpha}^{2} \quad , \; g = \nu/2q, \quad h = 2q^2/\nu \tag{9}$$

where q and ν are the estimated mean and variance, respectively, of the Q-statistic from the reference times. The Q-statistic determines the error between the predicted values using the NOC samples and the data arising from each chromatogram in the data series.

Hotelling's T^2 can be calculated as follows [12-14]:

$$D_i = \mathbf{t}_i^{\mathrm{T}} * \mathbf{S}^{-1} * \mathbf{t}_i \sim \frac{L(n^2-1)}{n(n-L)} F(L, n-L, \alpha) \tag{10}$$

where $\mathbf{S}$ is the variance-covariance matrix of the scores. The D-statistic follows the F-distribution with n and n-L degrees of freedom. For example, if α = 0.99, L = 3, and n = 27, the obtained F is 3.73. Multiplying by $\frac{L(n^2-1)}{n(n-L)}$(= 3.37) in the next step, the control limit was obtained as 12.56. The D-statistic indicates how close the chromatographic profile of the sample is to that of NOC region.

Contribution plots. It is possible to consider the contribution of each variable to the Q- and D-statistics, to distinguish the most responsible variables in the deviation of the process from NOC. The contributions from the jth variable in sample i for Q- and D-statistics can be obtained as follows:

$$\omega^{\mathrm{Q}}_{\mathrm{i,kj}}=(\mathrm{err}_{\mathrm{i,kj}})^2 \tag{11}$$

$$\omega_{i,jk}^{D} = \hat{\mathbf{t}}_i \left(\frac{\mathbf{T}'\mathbf{T}}{N-1}\right)^{-1} (x_{i,jk}\mathbf{p}_{kj}) \tag{12}$$

Plotting these values along the sample number for each of the three reactors, the contribution of each variable for each reactor and each sampling time point can be observed. If we plot the values for one sample, the contribution of each variable for that sample to Q- and D-statistics can be seen.

Unfolding MPCA

In reactor number (K)-unfolding-row-augmented MPCA (Fig. 1), the rows of the unfolded $\mathbf{X}$ matrix represent the sampling times. The model is expressed as:

$$\mathbf{X}_{I\times JK} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \tag{13}$$

where $\mathbf{T}$ is the score matrix of I×L, L is the rank, $\mathbf{P}$ is the loading matrix of (JK×L), and $\mathbf{E}$ is the residual matrix of I×JK [16]. For the reactor number (K)-unfolding-col-augmented MPCA the matrix $\mathbf{X}_{KI\times J}$ was obtained. Unfolding can also be performed in other directions, which result in IJ×K and/or J×KI unfolded matrices. As before, the next step is decomposition of the two-way data into score and loading matrices, using PCA.

Theoretically, compared to a bilinear model such as multiway principal component analysis (MPCA), a trilinear model such as PARAFAC, characterizes the relationship between the modes of variable and time, as opposed to simultaneously modelling all combinations of variables and

Archive of SID

www.SID.ir

times, *i.e.* it convolves the time and variable effects. Thus a trilinear model is comparatively simpler and more easily interpreted than a bilinear representation. In addition, for a trilinear model, fewer parameters are required to be estimated ($L \times(I+ J+ K)$ as opposed to $L \times(I+ J\times K)$). Conversely, the trilinear approach runs the risk of oversimplifying the model by assuming that an approximate trilinear structure exists, *i.e.* a similar relationship exists between variables through the entire batch. In contrast, for MPCA, correlation between arbitrary time-variable pairs can be modelled. It should also be noted that for PARAFAC the factors are not unique, *i.e.* calculating a different number of factors will result in a different set of loadings. This is in contrast to the bilinear techniques where the loadings are unique to a specific principal component, *i.e.* the models are nested.

Procedure

Sampling and dilution were the two steps utilized before high performance liquid chromatography (HPLC) for performing online analysis. To control sampling, dilution and HPLC parameters, a personal computer was used. Applying a proper sampling, a representative sample from the chemical process was drawn. The volume of sample pumped into the dilution device was 365 µl. The system diluted the sample into approximately 16.3 ml of ethyl acetate (Fisher Scientific, Loughborough, UK), to be injected into HPLC. Once diluted, the sample was pumped into an Agilent 1100 HPLC system controlled by Agilent ChemStation, v10.02 (Agilent Technologies, Stockport, Cheshire, UK). A micro-degasser, binary pump, column heater component and a variable wavelength UV detector (fitted with a standard 13 µl flow cell) were the components of the Agilent 1100 instrument. A 2.5 min reversed phase gradient method was programmed using acid-modified eluent pumped at 2 ml min^{-1} through a Zorbax-SB-C18 HPLC column (Agilent Technilogies), was controlled at 40 °C. The volume of sample injected into HPLC was 0.5 µl and chromatograms were recorded at 245 nm (single wavelength). To prevent carry-over from previous samples, the dilution device was programmed to flush itself with fresh solvent (diluent) between successive analyses. Over a period of 48.0 h 51, 177 and 186 samples were recorded from three reactors. The collected chromatographic data from Agilent ChemStation was exported to MATLAB v7.0 (Mathworks, Natick, MA, USA) for further analysis. All software was written in MATLAB.

RESULTS AND DISCUSSION

Overlaid plot of raw chromatograms of all 51 samples in the first reactor, all 177 samples in the second reactor and all 186 samples in the third reactor are shown in Figs. 3a to 3c. As it is illustrated in Fig. 3d, the baseline drift exists and should be corrected before further data analysis. Figure 3e includes the plots of chromatograms of the second reactor after pre-processing the data, including baseline correction and peak alignment. The only remaining point about the data was the different number of samples for each reactor, 51 for the first, 177 for the second, and 186 samples for the third one. To make the same number of samples in all three reactors, shape-preserving piecewise cubic Hermite interpolation was applied on the data and three 266×1981 chromatographic data were estimated for the three reactors. In this way, a three-way tensor of data $\underline{\mathbf{X}}$(266×1981×3) was obtained, with I = 266 samples in rows, J = 1981 retention times in columns, and K = 3 reactors in layers (depth).

To study the reactors simultaneously, a fixed size moving window factor analysis (FSMWFA) (using window width of 10 columns) was applied on the k-unfolded-row aligned data, $\mathbf{X}_{(I\times KJ)}$, to find the steady state region in the process. The ratios of the first evolutionary eigenvalues to the second ones at different windows are shown in Fig. 4. Samples with eigenvalue ratios higher than 500 show the region with one significant principal component and could be considered to be in the steady state. The values of this ratio are very small for the initial samples and some other regions which implies that the samples in these regions deviate more than the samples in the other regions from the steady state (Normal operating condition (NOC)). In this study, sample numbers 69 to 95 (27 samples) were selected as the NOC and the training set for MSPC.

PARAFAC and MSPC Charts

The next step was applying PARAFAC to the selected reference region, that is, samples 69 to 95. PARAFAC was applied using three significant factors. Vectorized matrix of loading values for the three significant factors in one of the

www.SID.ir

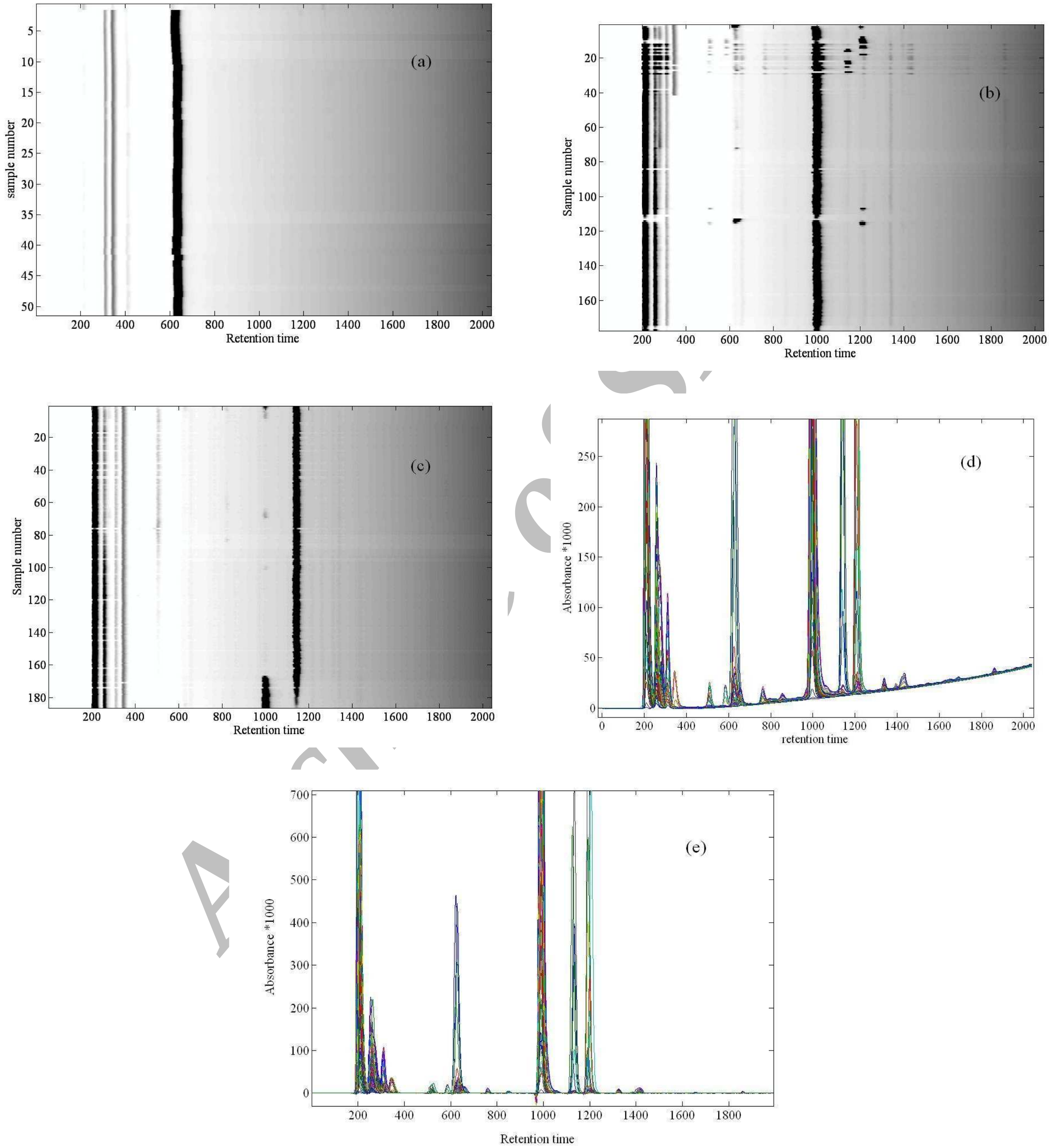

Fig. 3. Contour plots of overall chromatographic profiles of raw data from all 51 samples of the first reactor (a), 177 samples of the second reactor (b) and 186 samples of the third reactor (c). Plots of raw data (d) and baseline corrected and peak aligned data of 177 chromatograms of second reactor.

www.SID.ir

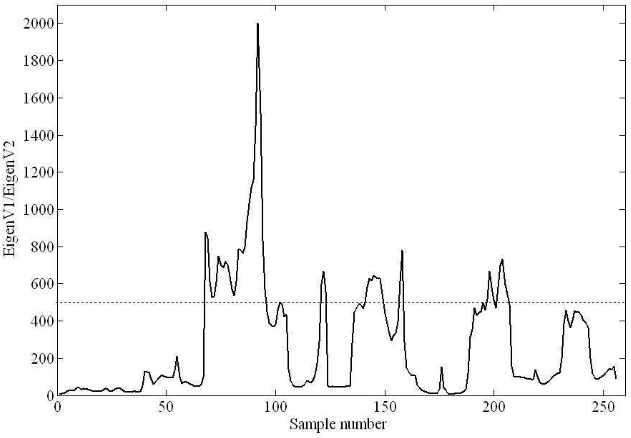

Fig. 4. Ratio of the first eigenvalues to the second eigenvalues obtained from application of FSMWEFA on the unfolded data of all 266 samples, for a window size of 10 datapoints.

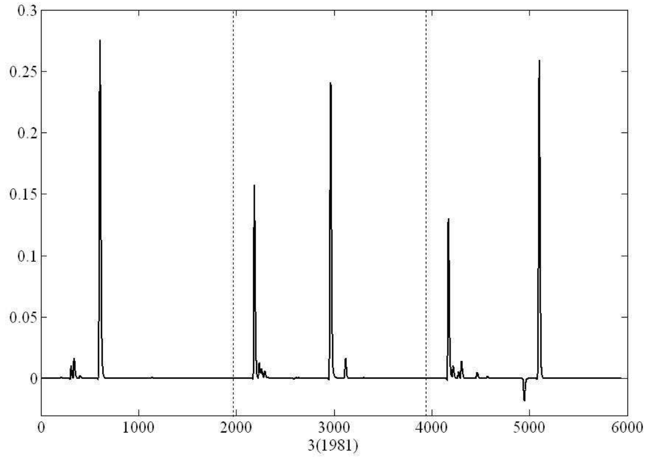

Fig. 5. Vectorized $\mathbf{B_r}$ loadings from application of PARAFAC on the reference NOC data.

Archive of SID

three modes shown in Fig. 5. The figure shows an average for the shape of chromatograms in each reactor, including the main peak, but not the peaks due to impurities. D-chart results for all 266 samples, based on 27 training set region, are illustrated in Fig. 6a. All samples in the NOC region are below 99% and 95% control limits. A number of initial samples are completely out of control, as they are from the start of the reaction before the products are formed. Samples from most of the regions are within the control limits. The Q-chart in Fig. 6b shows a number of samples between number 130 to 170 that are out of 95% limit and within 99% limit. The first 35 samples appear completely out of control, as they are from the start of the reaction before the products are formed, and thus are quite typical of the reaction. The other samples are within the limits. There is a region between samples 180 and 190 that is out of the limits in D chart and within the confidence limits in Q chart. This indicates that impurities are not serious in this region. However, the change in experimental conditions resulted in a higher values of D. For the regions 130 to 145 and 200 to 210, the reverse is true. The D values are below the limits and the Q statistic values are above the 95% limit which reveal the presence of impurities in this region. Samples from 100 to 120 are close to the NOC condition, considering both Q and D statistical parameters.

Figure 6c is the contribution plot for the D chart of all 266 samples from the three reactors. Peaks in this plot are the same as those in Fig. 5, which are due to the main components in the reaction. The fluctuation in the conditions for the first reactor is low, as shown by the small error peak about 600. Large peaks at about 2150 and 2950 show high fluctuations in experimental conditions for the second reactor. Unstable experimental conditions are mostly due to the initial times of reaction in which the reactor is not in the steady state. For the third reactor the condition is something between the 1st and 2nd reactors. Figure 6c shows the separate contribution of each of 266 samples in D-chart. For the regions with lower D values, such as samples 65 to 110 and 195 to 210 the dark bands are narrow, which show the lower contribution of the samples in D plot.

Contribution plot of Q-chart (Fig. 6d) includes some additional peaks, compared to Fig. 5. The additional peaks are about 2600, 3200 and 5000 and are due to the presence of impurities in the 2nd and 3rd reactors. The impurity peaks are due to samples with high values of Q-statistic, such as samples 240 to 250 and a number of initial samples. Results show that the Q contribution plot is sensitive to impurities. Thus, both the Q and D contribution plots are needed to detect the possible changes in the process, and provide complementary

www.SID.ir

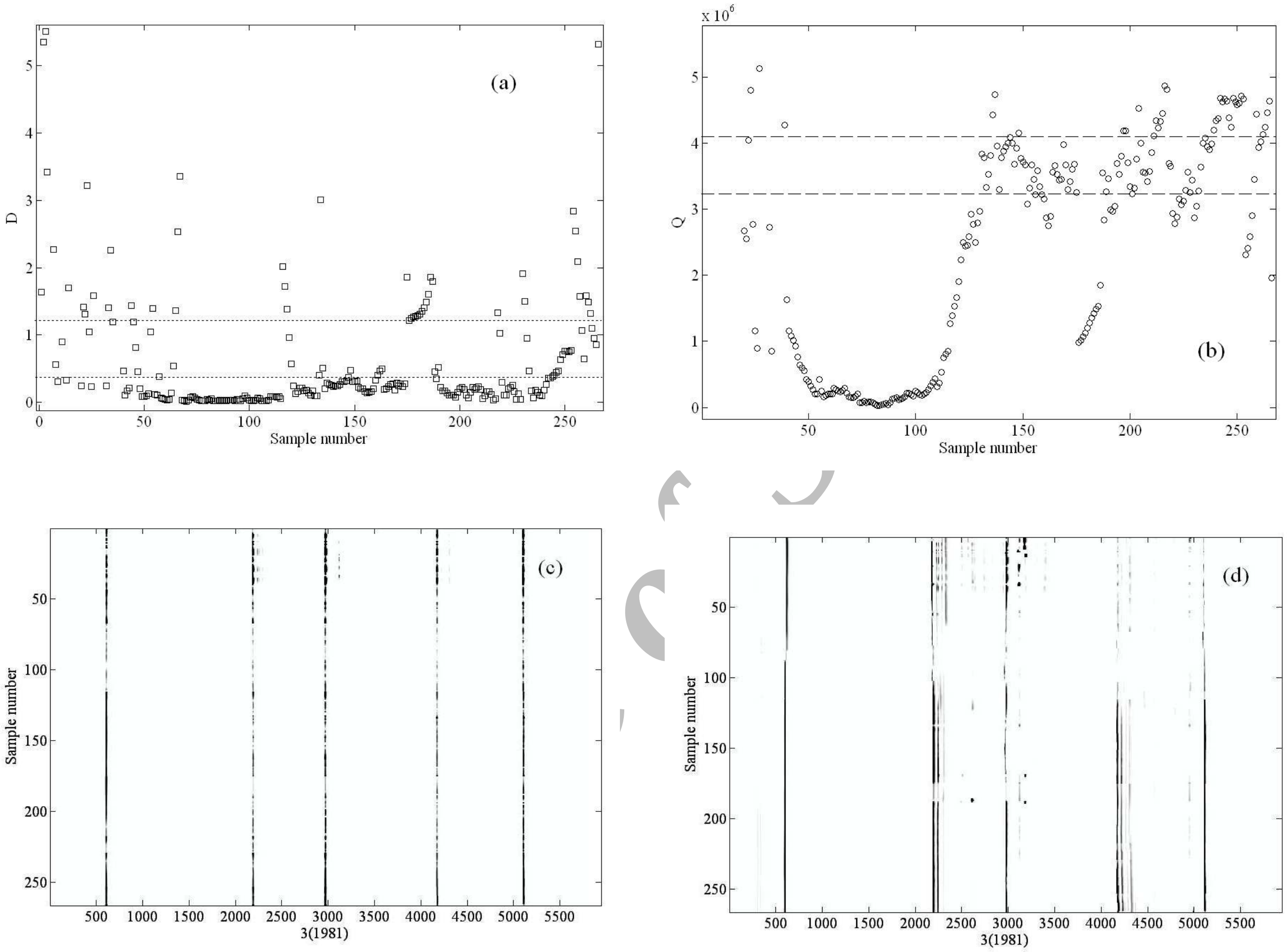

Fig. 6. D-chart (a) and Q-chart (b) for all 266 samples with 95% and 99% confidence limits. D contribution plot (c) and Q contribution plot (d) for all 266 samples.

information. Figure 6d shows the individual contribution of each sample in the Q-chart. The figure well shows that the impurities which are about 2600, and 3200 are due to samples 170, 190 and a number of initial samples in reactor 2.

Unfolding MPCA

The score and loading plots from reactor (*K*)-unfolding-column-augmented MPCA are shown in Figs. 7a and 7b. The first three principal components capture more than 90% of variations in the data. Scores and loadings that are assigned to a continuous line are due to the 1st reactor and show that the fluctuations are minmum among the reactors. The specific peak for the first reactor in the loading plot is related to the retention time 600. As it is shown in both score and loading plots, variations in the second and third reactors are linear combinations of the second (squares) and third (circles) factors, and two peaks, one about 1000 and the other between 1100 to 1200, contain the information about the second and third reactors, not selectively. The most intensive fluctuations belong to the second reactor, and specially to the initial times of the sampling. A sudden decrease in the most significant scores (squares) in the 2nd reactor (about abscissa value of 440

www.SID.ir

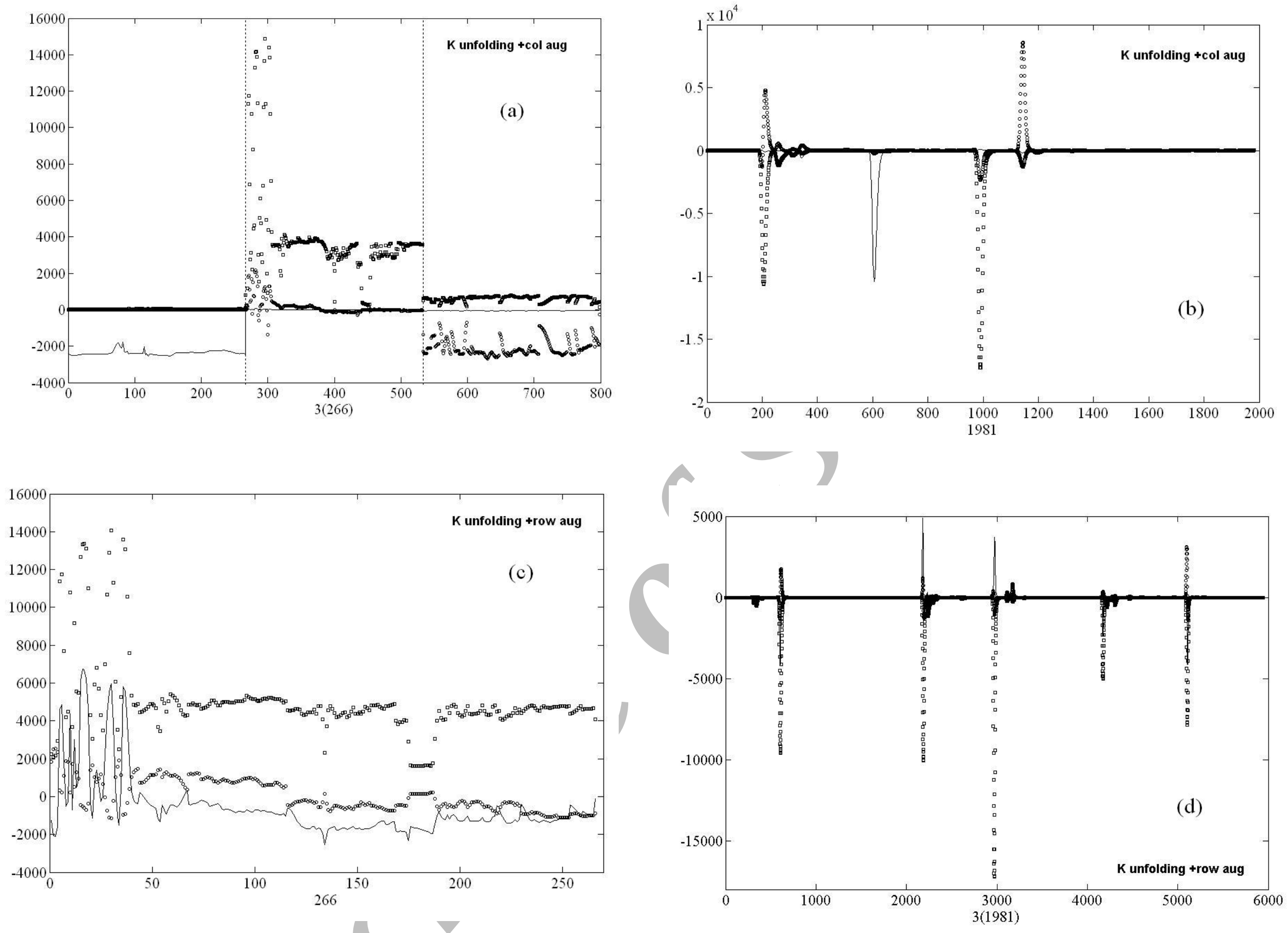

Fig. 7. Results from MPCA on reactor (*K*)-unfolded data. (a) and (b) are scores and loadings plots from decomposition of column-augmented data after unfolding. (c) and (d) are scores and loadings plots from decomposition of row-augmented data after unfolding.

in score plot, which means sample 445-266 = 179) is also indicated with high values of D-statistic (Fig. 6a).

The scores and loading plots from reactor (*K*)-unfolding-row-augmented MPCA are shown in Figs. 7c and 7d. In this case, too, more than 90% of variations is described by the first three principal components. Variations in all three reactors are described as a linear combination of the three principal components. Score plot shows that the first PC (squares) is more significant than the other two PCs, which shows that the considered process is a continuous one with the ideally steady state condition. Score plot also illustrates the fluctuations of the experimental conditions in the initial stages of the reaction as well as a sudden change in the experimental conditions for samples around 180 which accord with the D chart.

Scores and loading plots from sample (*I*)-unfolding-row-augmented MPCA on data are shown in Fig. 8. The information content of the plots are similar to score and loading plots from *K*-unfolding-column-augmented MPCA of the previous part. Three first PCs contain more than 90% of variations in the data. As can be understood from the loading plot (Fig. 8b) PC1 (square), PC2 (cross), and PC3 (circle) are due to the 1st, 2nd and 3rd reactors, respectively. Scores plot

www.SID.ir

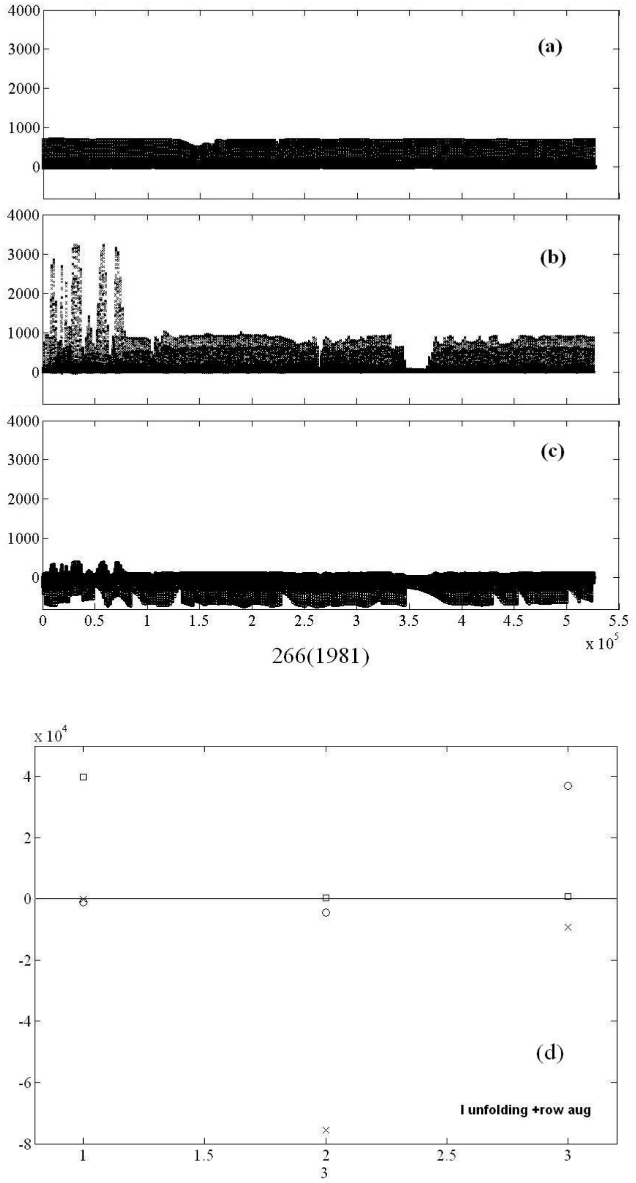

Fig. 8. Scores (a-c) and loadings (b) plots from decompositions of sample(*I*)-unfolded-row-augmented data.

shows that most fluctuations are related to the 2nd reactor and to the early stages of reactions. There is a sudden change in the scores in x = 360000 for reactors 2 and 3, which is probably due to the sample number 360000/1981 = 182. This sudden change is also illustrated in D-plot. The most stable reactor, from the starting times, is reactor 1.

CONCLUSIONS

This work is among a few studies which are based on using on-line HPLC for real-time monitoring of reactions associated with chemometrics technique. An initial treatment of data obtained from HPLC is necessary. The treatment includes baseline correction and alignment of peaks. The next step is applying MSPC on the treated data. The applied techniques are multivariate unfolding PCA and PARAFAC. PARAFAC method is combined with a Q-statistic and T^2-Hotelling's statistic. Techniques in addition to statistical tests show the deviation of the process from NOC in some regions. D-charts are more sensitive to fluctuations in the experimental conditions, whereas the Q-charts appear to be more sensitive to impurities or by-products.

Similar to D- and Q-charts, reactor-unfolding-row-augmented MPCA is more sensitive to overall variations during the investigation of the process, while reactor-unfolding-column-augmented MPCA and sample-unfolding-row-augmented MPCA are sensitive to variations in each reactor. PARAFAC method maintains an equal balance in terms of detecting both variations. Reactor 1 showed the least fluctuations in the conditions and was the best among the three reactors. This fact is illustrated in the score plots from MPA-based methods in addition to D- and Q-contribution plots, which are from PARAFAC.

Methods described in this report represent powerful tools for monitoring the reactions, and have potential to be extended to real-time applications. There is no clear agreement as to the most effective approach for the analysis of process data. The applied methods are complementary to each other and a well-trained chemometrician/practitioner should find both approaches to be useful for the continuous process data analysis. In many conditions, the observations and interpretations of the loading plots obtained from PARAFAC and the scores and loading plots from MPCA methods directly lead to similar findings.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support from the

of SID

www.SID.ir

Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran, and the Center for Chemometrics, the University of Bristol, UK.

REFERENCES

[1] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.

[2] R.G. Brereton, Analyst 11 (2000) 2125.

[3] R.G. Brereton, Chemometrics Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003.

[4] T.J. Thurston, R.G. Brereton, Analyst 5 (2002) 659.

[5] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206.

[6] A.R. Carvalho, J. Wattoom, L. Zhu, R.G. Brereton, Analyst 1 (2006) 90.

[7] P. Nomikos, J.F. MacGregor, Technometrics 37 (1995) 41.

[8] T. Kourti, J.F. MacGregor, Chemom. Intell. Lab. Syst. 28 (1995) 3.

[9] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Chemom. Intell. Lab. Syst. 51 (2000) 95.

[10] H. Hotelling, J. Educ. Psychol. 24 (1933) 417.

[11] M.J. Doyle, B.J. Newton, *CAST*, Dionex Corporation, 2002.

[12] B. Newton, K. Somerville, Water Tech. Conf., 1999.

[13] T.O. Passell, J. Chromatogr. A 671 (1994) 331.

[14] T.M. Larson, Proc. IFPAC/PAC Conf., 2001.

[15] J.C. Thompsen, Proc. Cont. Qual. 2 (1992) 55.

[16] P. Nomikos, J.F. MacGregor, AIChE J. 40 (1994) 1361.

[17] N. Gallangher, B.M. Wise, C.W. Stewart, Comput. Chem. Eng. 20 (1996) S739.

[18] S. Wold, N. Kettaneh, H. Friden, A. Holmberg, Chemom. Intell. Lab. Syst. 44 (1998) 331.

[19] H. Rannar, J.F. MacGregor, S. Wold, Chemom. Intell. Lab. Syst. 41 (1998) 73.

[20] B.M. Wise, N.B. Gallagher, S.W. Butler, J.E.D.D. White, G.G. Barna, J. Chemometr. 13 (1999) 379.

[21] D.J. Louwerse, A.K. Smilde, Chem. Eng. Sci. 55 (1999) 1225.

[22] R. Bro, Chemometr. Intell. Lab. Syst. 38 (1997) 149.

[23] A.K. Smilde, J. Chemometr. 15 (2001) 19.

[24] A.K. Smilde, R. Bro, P. Geladi, Multi-Way Analysis: Applications in the Chemical Sciences, John Wiley & Sons, 2004.

[25] R. Leardi, C. Armanino, S. Lanteri, L. Alberotanza, J. Chemometr. 14 (2000) 187.

[26] A.K. Smilde, Chemom. Intell. Lab. Syst. 15 (1992) 143.

[27] A.J. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemometr. 13 (1999) 397.

[28] D.J. Louwerse, A.K. Smilde, Chem. Eng. Sci. 55 (2000) 1225.

[29] B. Lennox, G.A. Montague, H. Hiden, G. Kornfeld, P.R. Goulding, Biotechnol. Bioeng. 74 (2001) 125.

[30] A. Simoglou, E.B. Martin, A.J. Morris, Comp. Chem. Eng. 26 (2002) 909.

[31] J. Lee, C. Yoo, I. Lee, Comp. Chem. Eng. 28 (2004) 1837.

[32] C.K. Yoo, J. Lee, P.A. Vanrolleghema, I. Lee, Chemometr. Intell. Lab. Syst. 71 (2004) 151.

[33] N. Lu, Y. Yang, F. Gao, F. Wang, in: F. Allgower, F. Gao (Eds.), Proceedings of 7th International Symposium on Advanced Control of Chemical Processes, 2004, pp. 471-476.

[34] B.M. Wise, N.L. Ricker, Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity. IFAC Symp. on Advanced Control of Chemical Processes, Toulouse, 1991, p. 125.

[35] P. Miller, S.E. Swanson, C.F. Heckler, Int. J. Appl. Math. Comput. Sci. 8 (1998) 775.

[36] B.M. Wise, N. Gallagher, E.B. Martin, J. Chemometr. 15 (2001) 285.

[37] F.N. Fritsch, R.E. Carlson, SIAM J. Numer. Anal. 17 (1980) 238.

[38] D. Kahaner, C. Moler, S. Nash, Numerical Methods and Software, Prentice Hall, Englewood Cliffs, NJ, 1988.

[39] N.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.

[40] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231.

[41] S.K. Setarehdan, J. Chemometr. 18 (2004) 414.

[42] Z.-D. Zeng, C.-J. Xu, Y.-Z. Liang, B.-Y. Li, Chemom. Intell. Lab. Syst. 69 (2003) 89.

[43] A.K. Elbergali, R.G. Brereton, Chemom. Intell. Lab. Syst. 23 (1994) 97.

Archive of SID

www.SID.ir

[44] F. Cuesta Sanchez, B.G.M. Vandeginste, T.M. Hancewicz, D.L. Massart, Anal. Chem. 69 (1997) 1477.

[45] H.R. Keller, D.L. Massart, Anal. Chim. Acta 246 (1991) 379.

[46] H.R. Keller, D.L. Massart, Y.Z. Liang, O.M. Kvalheim, Anal. Chim. Acta 263 (1992) 29.

[47] R.B. Cattell, Psychometrika 9 (1944) 267.

[48] J.D. Carroll, J.J. Chang, Psychometrika 35 (1970) 283.

[49] R.A. Harshman, UCLA Working Papers Phonet. 16 (1970) 1.

[50] R. Bro, Multi-Way Analysis in the Food Industry: Models, Algorithms, and Applications. PhD Thesis, Royal Veternary and Agricultural University, Fredericksberg, 1998.

Archive of SID

www.SID.ir