Journal of Chromatography A, 1217 (2010) 6503–6510

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Chemometric quality control of chromatographic purity

Kristoffer Laursen[a,b,*], Søren Søndergaard Frederiksen[b], Casper Leuenhagen[b], Rasmus Bro[a]

[a] Quality & Technology, Department of Food Science, Faculty of Life Sciences – University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark
[b] Novo Nordisk A/S, 2880 Bagsværd, Denmark

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form 12 August 2010
Accepted 16 August 2010
Available online 21 August 2010

Keywords:
Chromatographic pattern monitoring
Impurity detection
Overlapping peaks
Principal component analysis (PCA)
Multivariate statistical process control (MSPC)
Signal preprocessing

ABSTRACT

It is common practice in chromatographic purity analysis of pharmaceutical manufacturing processes to assess the quality of peak integration combined by visual investigation of the chromatogram. This traditional method of visual chromatographic comparison is simple, but is very subjective, laborious and seldom very quantitative. For high-purity drugs it would be particularly difficult to detect the occurrence of an unknown impurity co-eluting with the target compound, which is present in excess compared to any impurity. We hypothesize that this can be achieved through Multivariate Statistical Process Control (MSPC) based on principal component analysis (PCA) modeling. In order to obtain the lowest detection limit, different chromatographic data preprocessing methods such as time alignment, baseline correction and scaling are applied. Historical high performance liquid chromatography (HPLC) chromatograms from a biopharmaceutical in-process analysis are used to build a normal operation condition (NOC) PCA model. Chromatograms added simulated 0.1% impurities with varied resolutions are exposed to the NOC model and monitored with MSPC charts. This study demonstrates that MSPC based on PCA applied on chromatographic purity analysis is a powerful tool for monitoring subtle changes in the chromatographic pattern, providing clear diagnostics of subtly deviating chromatograms. The procedure described in this study can be implemented and operated as the HPLC analysis runs according to the process analytical technology (PAT) concept aiming for real-time release.

1. Introduction

Product purity is of utmost importance in ensuring drug quality; consequently, impurities must be monitored carefully. In general, impurities present in excess of 0.1% relative to the target compound in drug substances should be detected and identified as by the ICH requirements [1]. Analytical separation techniques based on high performance liquid chromatography (HPLC) are commonly used for purity analysis in biopharmaceutical manufacturing processes. The separation and subsequent detection of compounds in a sample delivers a chromatogram, which ideally allows to identify individual peaks and to attribute them to individual compounds. Typical purity analysis in industrial processes usually deals with a manageable amount of well known peaks of compounds at relatively high concentrations. This can easily be handled automatically with available software packages suitable for routine analysis of chromatograms [2]. However, generic peak detection algorithms may often suffer from inconsistent reliability towards unknown peaks with low signal to noise ratio and overlapping peaks of different shapes. Thus, it is common practice to assess the results of peak integration by visual inspection of the chromatogram. Visual inspection of chromatograms has been used for decades [3] and is a valid procedure for identification of protein samples recognized by the regulatory authorities [4,5]. Although simple, this partly manually method is quite laborious, extremely time consuming, seldom quantitative and prone to subjective decision-making probably causing additional errors. To comply with increased focus on process analytical technology (PAT) and quality by design (QbD) (aiming for enhanced process understanding that improves process control moving towards continuous quality verification and real-time release of an end product) there is a need for an automatic and timely tool for objectively monitoring the chromatographic pattern. Even though various advanced approaches have been published towards automatic peak detection [2,6,7], there still is a need for a tool to detect relevant subtle differences in the chromatographic pattern both quantitatively and in a statistically reliable way.

New impurities mainly originate during the synthesis process from raw materials, solvents, intermediates, and by-products [8]. For high-purity drugs, the target compound is present in excess compared to any impurity. Hence, occurrence of an unanticipated impurity co-eluting with the target compound is a particular problematic challenge. In such cases, it would be difficult or impossible

* Corresponding author at: Novo Nordisk A/S, 2880 Bagsværd, Denmark. Tel.: +45 30795458.
E-mail address: krfl@novonordisk.com (K. Laursen).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.08.040

to spot the impurity peak visually and the peak integration may therefore not be able to identify and separate impurity and target peaks. Commercially available chromatographic pattern matching software has been studied to differentiate whole chromatograms objectively and quantitatively [9]. Such pattern matching analysis tool compares chromatograms in pairs, where one is specified as reference. However, in most processes it would be a difficult task to identify one representative reference chromatogram. As a result, several chromatograms representing common-cause variation should be included for reference. This can be achieved with multivariate statistical process control (MSPC) based on latent variable methods such as principal component analysis (PCA) [10,11]. MSPC based on latent variable methods have been used over the last 20 years and has revolutionized the idea of statistical process control for multivariate purposes [12]. The entire chromatogram can be monitored by the operator looking at only a few multivariate control charts, which are simple and easy to understand. MSPC based on PCA has previously been applied on integrated peak tables derived from chromatographic data and proven as a valuable tool to compare chromatograms [7,13]. This approach is valid when peaks are clearly unimodal (one maximum only). Such an approach cannot handle embedded- or non-resolved peaks, which consequently would be integrated as one peak. The unimodality assumption is most often far from reality, and therefore inclusion of as much chromatographic information as possible is wanted when applying PCA. So far, MSPC based on PCA applied directly on raw chromatograms has not yet been reported. With such a technique historical chromatograms can be exploited for empirical modeling to monitor and diagnose subtle changes in future chromatographic patterns. Nevertheless, multivariate data analysis using the raw chromatogram as input data is very sensitive to chromatographic *artifacts* such as baseline- and retention time drift [14]. Therefore, mathematical preprocessing of chromatograms is a crucial step in order to generate as clean data as possible. In addition, it may be necessary to preprocess the clean data further in order to emphasize the relevant (chemical) information before PCA is applied [15].

In this study, we develop and investigate the sensitivity of MSPC based on PCA for monitoring, detection and diagnosis of small and embedded impurity peaks appearing in analytical chromatography. The case study considers historical HPLC chromatograms from biopharmaceutical in-process analysis of a high-purity drug substance.

2. Theory and methods

The development of a method for chemometric quality control of chromatographic purity follows a modified version of a previously described trajectory [16]. The trajectory is divided in three phases; the initial phase, the training phase and the application phase (ITA) as illustrated in Fig. 1.

In the initial phase, appropriate historical chromatograms are collected and prepared for PCA modeling. In the training phase a PCA model based on normal operation condition (NOC) chromatograms is developed (describing common-cause variation) and MSPC charts are constructed. Finally, in the application phase new chromatograms are fitted to the model and monitored using the control charts developed in the training phase. Deviating chromatograms are diagnosed using contribution plots to determine causes of the deviating behavior.

2.1. Signal preprocessing

The variation in chromatograms from an HPLC analysis is the sum of uninduced- and induced variations. The uninduced variation is all the variation originating from uninduced chemical variance, sampling, sample work-up, and analytical variation. The most significant uninduced variation in chromatography is baseline- and peak drift. Novel and advanced signal preprocessing algorithms can be applied to handle these artifacts in order to obtain data appropriate for subsequent data analysis. Moreover, it may be important to scale the data before starting the chemometric analysis. Hereby, the aim is to focus on the induced variation and emphasize the chemical relevant information in the samples.

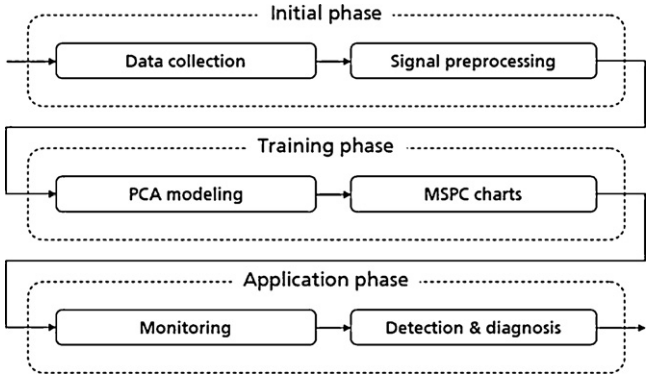

Fig. 1. The three phases according to ITA trajectory (initial, training and application phase).

2.1.1. Baseline correction

Baseline correction in chromatography is commonly employed to eliminate interferences due to baseline drift. Several baseline correction methods are available in the literature [17,18]. One efficient way of baseline correction operates in local regions of the chromatogram and uses B-splines constructed from polynomial pieces joined at certain positions (knots) [19]. The method operates by gradually eliminating points in the signal furthest (northern distance) away from the fitted polynomial until the number of selected support points (baseline points) is reached. Since the method works in local regions it is required that the number of knots and their position are set. This is actually an advantage as local changes in baseline can be corrected by placing more knots in the problematic regions. The method also requires input for the order of the polynomial that is fitted between the knots. Upon selecting the baseline-algorithm and its settings from initial data investigation, baseline correction can be an objective and automatic preprocessing.

2.1.2. Alignment

Alignment of shifted peaks can be performed in various ways. Very reproducible chromatographic data often need only a movement of the whole chromatogram a certain integer sideways for proper alignment. This is characterized by a systematic or linear shift and can easily be handled by the correlation optimized shifting (*coshift*) algorithm [20] or the recently published *icoshift* algorithm [21]. Yet, if the column is changed between runs or if samples are measured over a long period of time, more complex shift correction is needed. This non-systematic or non-linear shift is characterized by a different degree of shifts for multiple peaks across samples and can be seen as peaks shifting independently from one another in the same chromatogram. One effective method, which can handle non-systematic shifts in chromatographic data, is the piecewise alignment algorithm correlation optimized warping (COW) [22,23]. Both Coshift and COW algorithms align each chromatogram towards a target. The choice of a target chromatogram is an important aspect of the alignment methods considered here. Several methods for how to find a proper reference chromatogram can be used. Among these are, the average chromatogram, the first loading of a PCA model, the most inter-

similar chromatogram among all chromatograms or the sample run in the middle of a sequence. However, the choice depends on the homogeneity of the samples, on the degree of missing peaks across the chromatograms and many other things, which should be considered in each individual application [24,25].

2.1.3. Scaling

The choice of preprocessing procedure is crucial for performance of the subsequent chemometric analysis. For instance a 1000-fold difference in concentration for the target compound and an impurity is not proportional to the chemical relevance of these compounds [15]. Thus, an appropriate preprocessing may increase the sensitivity on detecting small impurity peaks hidden under the target peak by chemometric analysis and MSPC. Scaling methods are data preprocessing approaches that divide variables by a factor, which is different for each variable. The aim is to adjust for the disparity in fold differences between various signals by converting the data into differences in concentration relative to the scaling factor. One effective way to reduce the relative importance of large values without blowing up noise is square root mean scaling. This scaling method uses the square root of the mean (of individual variables) as scaling factor.

2.2. MSPC based on PCA

The goal of any statistical process control (SPC) scheme is to monitor the performance of a process over time. Most SPC schemes currently in practice are based on charting a single or a small number of product quality variables in a univariate way. This approach is inadequate for processes where massive amounts of highly correlated variables are being collected as is the case in chromatograms.

Latent variable methods such as PCA that handle all the variables simultaneously are required in these data-rich applications. PCA has previously proven a valuable tool to objectively compare entire chromatograms [26]. With PCA the information from many correlated variables in a chromatographic data matrix $X\,(M \times N)$ can be projected down onto a low-dimensional subspace defined by a few latent variables or principal components TP' and a residual part E $(M \times N)$:

$$X = TP' + E \tag{1}$$

where $T\,(M \times A)$ is the orthogonal score matrix and $P\,(N \times A)$ is the orthonormal loading matrix. The chromatographic pattern is then monitored in this A-dimensional subspace by using a few multivariate control charts built from multivariate statistics. Using the information contained in all the measured chromatographic variables simultaneously, these MSPC charts are much more powerful in detecting faulty conditions than conventional single variable SPC charts [27]. Once the MSPC chart signals a faulty alarm, the model can be scrutinized to understand the cause of the alarm; hereafter a possible corrective action can be taken. Variables responsible for the faulty signal, due to a disturbance in any of the subspaces can be projected back to the original variables and thereby identified. In general, there exist two ways to investigate the nature of the fault that causes the control chart to signal [28,29]. Faults that obey the correlation structure, but have an abnormal variation (i.e. extreme variation within the model) are described by the scores in Hotelling's T^2 also referred to as D-statistic. Hotelling [30] introduced the T^2 for principal components:

$$T^2 = \sum_{r=1}^{R} \frac{t_r^2}{\sigma_{t_r}^2} \tag{2}$$

where t_r is the rth principal component score, $\sigma_{t_r}^2$ is the variance of t_r and R denote the number of principal components retained in the PCA model. The D-statistic can be expected to approximately follow an F distribution and the confidence limits for the control chart can be calculated according to Jackson [31].

Faults that break the correlation structure (i.e. variation to the model) are represented in the sum of squared residuals also referred to as Q-statistic:

$$Q = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2 \tag{3}$$

where x_n and $\hat{x}_n$ are a measurement of the nth variable and its predicted (reconstructed) value, respectively. N denotes the number of process variables. Several ways to determine the confidence limits for the Q-statistic is described [32,33]. In the present paper, a normal distribution to approximate a weighted chi-square distribution is used from which the confidence limits for the Q chart can be calculated according to Jackson and Mudholkar [34].

Most commonly 95% or 99% confidence limits are used for both the D- and Q-statistics to determine whether a sample is considered an outlier. In the application described here a 99.73% confidence limit ($\sim 3\sigma$) is used as the upper control limit (UCL) similar to ordinary Shewart control charts. From the D- and Q-statistics, two complementary multivariate control charts are constructed. Chromatographic fault detection in the D-statistics could for example be caused by an increased load on the analytical column leading to intensified signals, but intact correlation between the chromatographic signals. If necessary, this load-effect may however be handled using normalization as preprocessing. Fault detection in the Q-statistics could for example be induced by the presence of a new peak in the chromatogram resulting in broken correlation between the chromatographic signals exemplified in Fig. 2. The sensitivity of fault detection towards changes in the chromatogram depends on the historical NOC data, chromatographic retention time window, preprocessing methods, and number of components included in the NOC model. If a new chromatogram falls outside the UCL in the D- or Q-statistics control chart, it is characterized as a fault and the chromatogram is considered to deviate significantly from the chromatograms included in the PCA model. It is not only important to detect that the chromatographic pattern is deviating, it is also important to search for the original chromatographic signals responsible for the fault. One of the most widely used approaches is using contribution plots [35–37]. Contribution plots compute a list of each single chromatographic signal (retention time) that contribute numerically to the D- and Q-statistics respectively. However, contribution plots do not reveal the actual cause of the fault. Therefore, those variables responsible for the faulty signal should be investigated, and incorporation of chemical and technical process knowledge may be necessary to diagnose the problem and discover the root causes of the fault [27]. As an enhancement to the way the faults are typically detected and source determined, it is possible to calculate confidence intervals for the residuals of individual variables, rather than only the overall residual [38].

2.3. Chromatographic simulation

The goal of chromatography is to separate different components from a solution mixture. The resolution expresses the extent of separation between the components in a sample, and is a useful measure of the columns separation properties of that particular sample. The higher the resolution of the peaks in the chromatogram, the better extent of separation between the components the column provides. A simplified method to calculate the resolution of a chromatogram is to use the plate model [39]. The plate model assumes that the column can be divided into a certain number of plates, and the mass balance can be calculated for each individual plate. This approach approximates a typical chro-

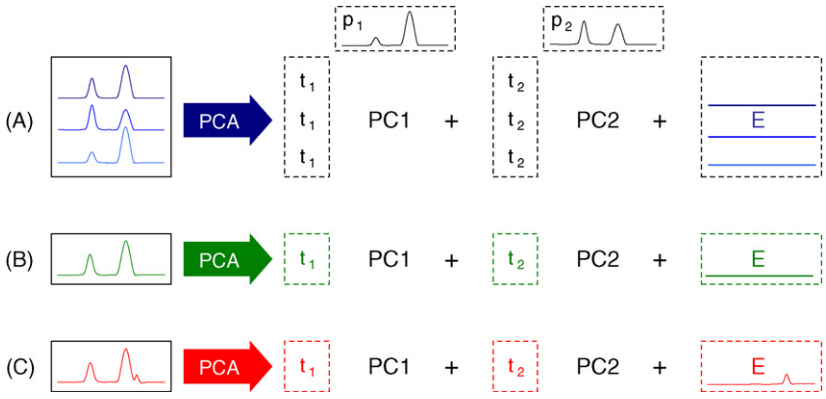

Fig. 2. Example of chromatographic pattern monitoring using PCA. (A) PCA modelling on NOC chromatograms using two principal components. (B) Prediction of a new chromatogram within common-cause variation. (C) Prediction of a new chromatogram deviating from common-cause variation resulting in abnormal residuals.

matogram curve as a Gaussian distribution curve. By doing this, the curve width is estimated as four times the standard deviation of the curve (4σ). Sigma can be estimated by calculating the segment of the peak base (w_b) intercepted by the tangents drawn to the inflection points on either side of the peak. The inflection points can be found by calculating max and min of the first derivative chromatogram [40]. The parameter σ is calculated as w_b divided by four. To define to what extent an impurity is hidden under the target peak; the peak resolution (R_s) is used [7]. R_s expresses the efficiency of separation of two peaks in terms of their average peak width at base [40]:

$$R_s = 2\frac{(t_{R2} - t_{R1})}{(w_{b1} + w_{b2})} \quad (4)$$

where t_{R1} and t_{R2} are the retention time of solute 1 and 2 respectively ($t_{R2} > t_{R1}$) and w_{b1} and w_{b2} are the Gaussian curve width of solute 1 and 2 respectively (the retention time is the time from the start of signal detection to the time of the peak height of the Gaussian curve). Usually, in chromatography the plate number is approximately constant for similar components with similar retention times. The plate number N for a Gaussian peak is given by [40]:

$$N = \left(\frac{t_R}{\sigma}\right)^2 \quad (5)$$

With similar retention times and plate numbers the peak width of the impurity and the target component is hence similar and a reasonable assumption is [40]:

$$R_s \approx \frac{t_{R2} - t_{R1}}{w_{b2}} \quad (6)$$

Based on these assumptions an impurity peak was generated as a pure Gaussian peak using σ calculated from the target peak in a randomly chosen chromatogram from the validation sample set. The generated impurity was subsequently added to the validated chromatogram. As mentioned previously, impurities present in excess of 0.1% relative to the target compound should be identified. Therefore, the relative amount of simulated impurity was kept constant at 0.1%. To give different degrees of chromatographic similarity between the target compound and the related impurity, the resolution (R_s) was varied from 0 (completely hidden) to 2 (well separated).

3. Experimental

Fifty in-process samples of a high-purity drug substance were collected for routine quality control testing. All samples were collected under NOC, i.e. the process has been running consistently and only high quality products have been obtained. The 50 samples represent a substantial time period so as to represent possible physical changes in the chromatographic system as well as changes in production arising e.g. from different batches of raw materials being used. The purity, measured by reverse-phase high performance liquid chromatography (RP-HPLC), was performed on a Waters Alliance HPLC system that consists of a Waters 2690 Separation Module (combined pump and autosampler) and a Waters 2487 Dual-Wavelength UV detector (Waters, Milford, MA, USA). The detection wavelength was 214 nm. The separation was performed on a reverse phase 125 mm × 4 mm i.d. 5 μm 100 Å column (FeF Chemicals, Køge, Denmark) by employing an isocratic elution followed by gradient elution. The mobile phase consisted of Eluent A (10%, v/v acetonitrile in sulphate buffer pH 2.5) and Eluent B (60%, v/v acetonitrile in water). Chromatographic data was collected using Empower 2 (Waters) and exported as the raw signals vs. time (ASCII/ARW files) to Matlab version 7 (Matworks, Natick, MA, USA) for further analysis. All software was written in Matlab using tools from PLS_Toolbox.

4. Results and discussion

4.1. Initial phase

The main goal of the training phase is to collect and prepare historical NOC chromatograms for modeling. Fifty historical HPLC chromatograms obtained for purity analysis of an industrial high-purity drug substance were collected and imported into MATLAB. The chromatograms were organized as an $M \times N$ data matrix X, with M rows or samples and N columns or elution times. A relevant chromatographic retention time window was chosen around the target peak, resulting in a 50 (samples) × 1500 (retention times) dataset/matrix. Coshift alignment was applied to handle larger systematic retention time shifts, followed by COW to handle non-systematic retention time shifts. Both algorithms align the chromatograms towards a manually chosen inter-similar target chromatogram as illustrated in Fig. 3 The use of both alignment methods clearly handles all the retention time shifts and delivers adequate aligned chromatographic profiles.

To reduce baseline drift, baseline-spline was applied to the dataset. In this case study a first order polynomial was chosen and 3 knots were positioned at retention time point 200, 1100 and 1300 (not shown).

To increase the sensitivity on detecting small impurities hidden under the target peak different centering, scaling and transformation methods were tested. Among these are mean centering, autoscaling, parato scaling, vast scaling, square root mean scaling, and log transformation. Most of the methods are described

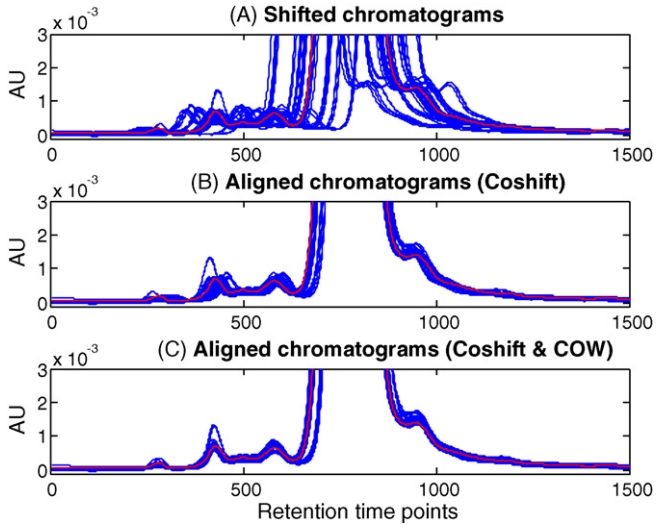

Fig. 3. Plot of shifted (A) and aligned (B and C) chromatograms (blue) towards a reference (red) using Coshift- and COW algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

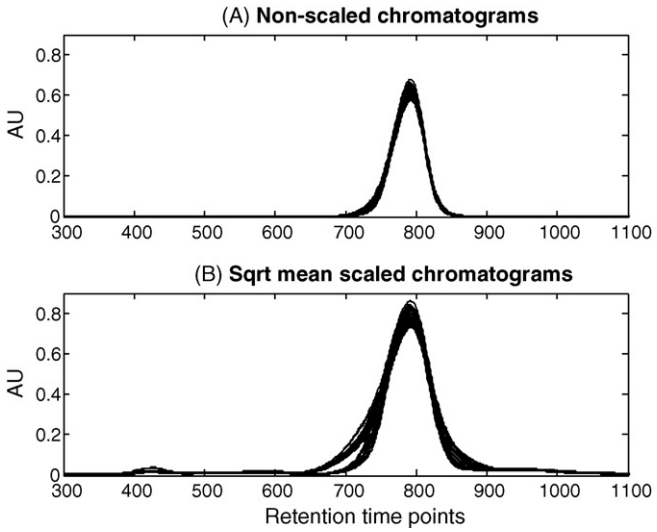

Fig. 4. Plot of chromatograms before (A) and after (B) square root mean scaling.

by [15,20]. The application of different preprocessing methods had very different effects on the resulting data (not shown). The methods were evaluated both by visual inspection of the resulting data and on the results obtained when used as input for subsequent data analysis in the training- and application phase. Square root mean scaling turned out to be the most appropriate preprocessing method for this particular application, as it first of all manages to adjust for the variation in fold differences between the target peak and the minor surrounding peaks without blowing up noise. Secondly, the characteristic appearance of the chromatogram is kept intact, which in this case is helpful when interpreting the contribution plot during the application phase. The result of square root mean scaling applied to the data is illustrated in Fig. 4.

4.2. Training phase

The essence of the training phase is to model the common-cause variation present in the chromatograms obtained under NOC. Since this NOC model exclusively determines whether a new chromatogram is similar or deviates significantly from the NOC chromatograms, the monitoring performance depends very much upon adequacy and representativity of these NOC chromatograms. The number of samples needed to construct a NOC model and control charts depends on the application. In this case study a calibration set consisting of the first 40 chronologically

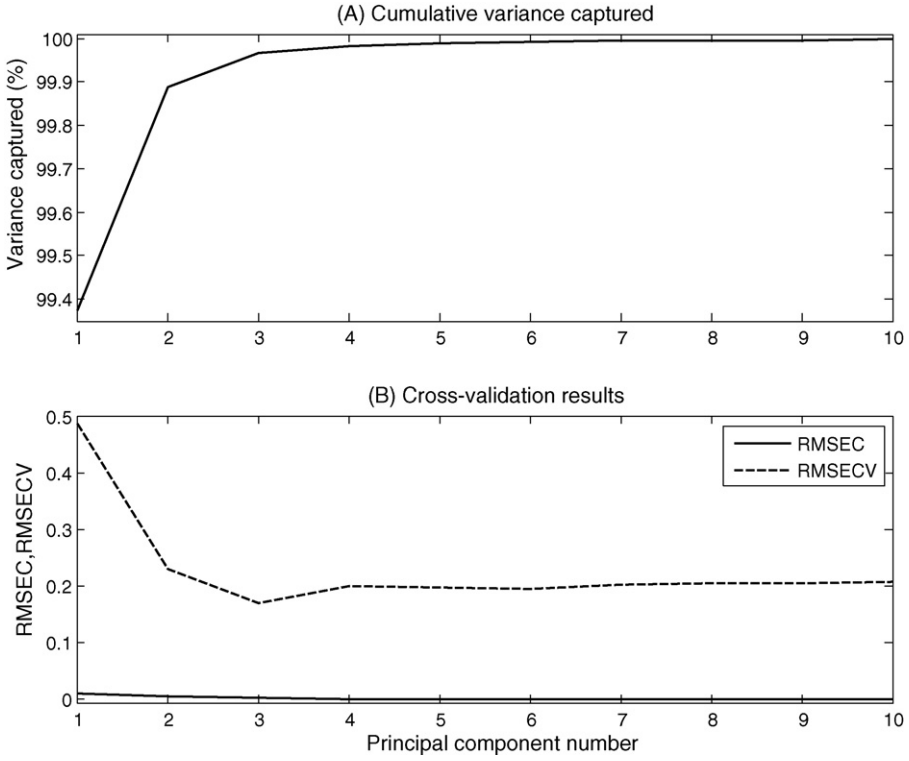

Fig. 5. Plot of cumulative variance captured (A) and results of leave-one-out cross-validation (B).

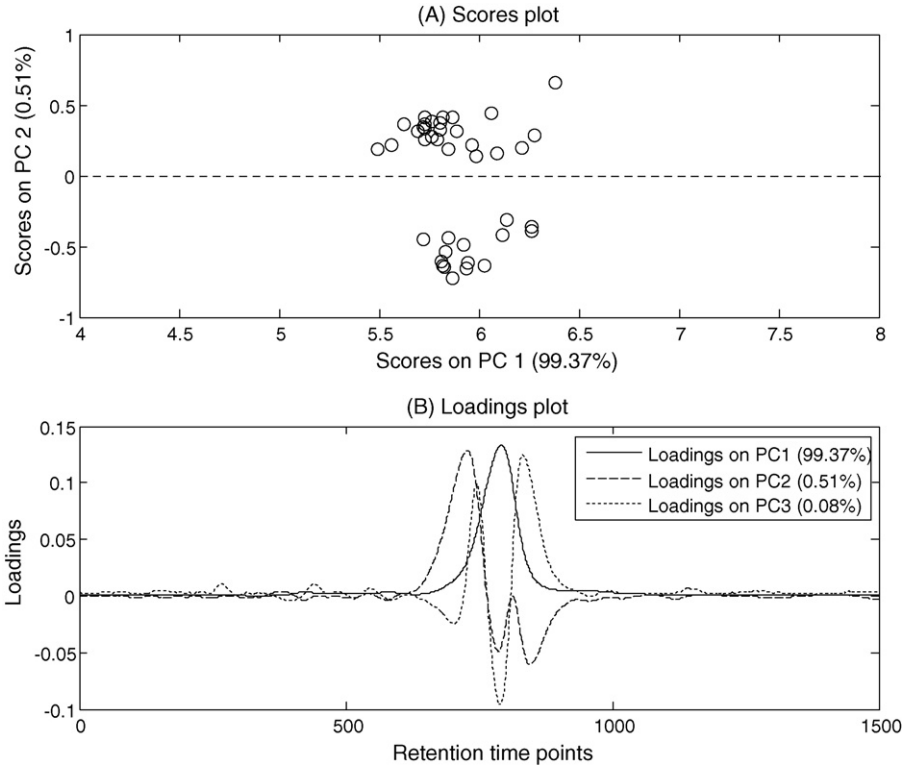

Fig. 6. Scores plot of PC2 vs. PC1 (A) and loadings plot on first three principal components (B).

ordered chromatograms was used to develop a three component PCA model describing 99.97% of the common-cause variation. We have selected an optimal number of three components based on the variance captured (Fig. 5a) and on the results of leave-one-out cross-validation (Fig. 5b). Both variance captured and root mean squared error of calibration (RMSEC) flattens out after three components, also root mean squared error of cross-validation (RMSECV) has the first clear local minimum at three components, indicating that after this point, the components just reflect noise. In addition, the inspection of loadings confirmed that only the first three components reflect real chromatographic variation (Fig. 6b). As the principal components higher than three are very noisy and do not seem to contain any clear systematic structure, it is appropriate to consider them as reflecting noise. Inspection of the scores plot provided in Fig. 6a showing PC2 vs. PC1, reveal that the calibration samples are separated in two groups in PC2. The corresponding loading for PC2 (Fig. 6b) indicated that this was due to an increased fronting and partly decreased tailing on the target peak. This chromatographic difference between the two groups of calibration samples most likely originate from analytical variation (ex. column, solvents, pump, temperature) not handled by the preprocessing. This chromatographic variation is also observed in Fig. 4b. However, no systematic pattern was recognized when plotting PC2 scores vs. chronologically ordered sample number (data not shown), which lead to the conclusion that the grouping observed in PC2 represents common-cause-variation. The model was validated using an independent validation set consisting of the last 10 chronologically ordered chromatograms. In Fig. 7 D- and Q-statistics of calibration and validation samples are presented with 95%, 99% and 99.73% (UCL) confidence limits.

By inspection of the D- and Q-statistics it can be confirmed that three components describe the common-cause variation (Fig. 7). All 50 NOC samples are within the 95% confidence interval in the D-statistic chart, whereas in the Q-statistic chart two samples (~5%) are outside the 95% confidence interval as expected from a normal distribution point of view. Both D- and Q-statistics are monitored during the training phase. Nevertheless, as this study focuses on purity analysis; we are primarily interested in the residuals. We use the residuals to identify new, unanticipated peaks, which are not part of the normal chromatographic pattern and thus, the model. On

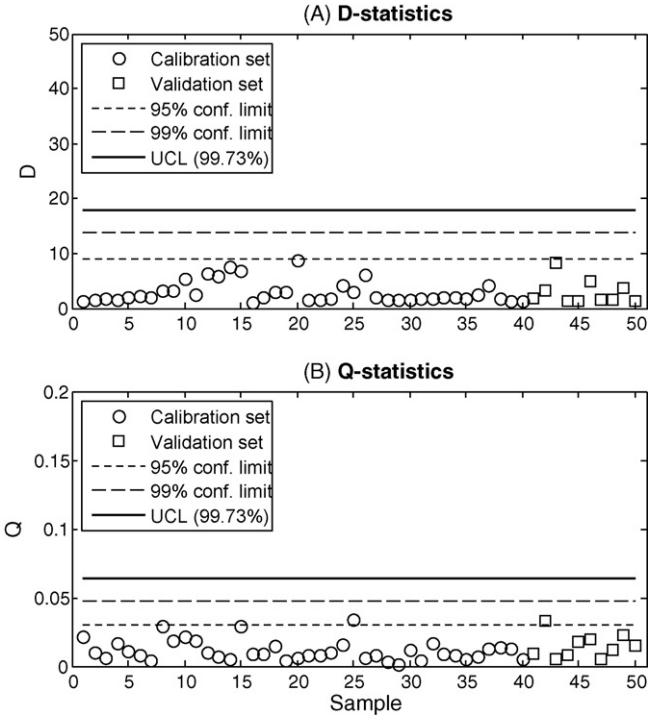

Fig. 7. Plot of (A) D-statistics and (B) Q-statistics of calibration (circle) and validation (square) sample sets.

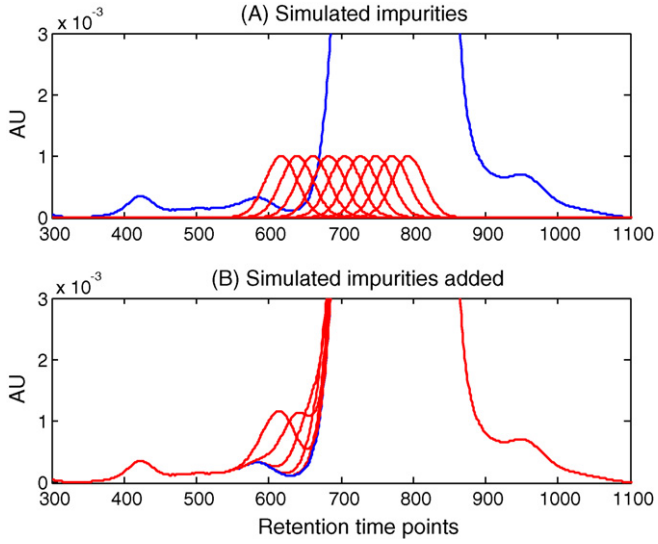

Fig. 8. Simulated 0.1% area impurity peaks (red) in 9 varied resolutions from 0 to 2 before (A) and after (B) added to a reference chromatogram (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

the other hand, when developing the model in the training phase, both the D- and Q-statistics are of interest. These statistics may contribute with important and complementary indications about samples to exclude from the NOC model as they do not describe common-cause variation and magnitude. In this case all 50 samples used in the training phase are within their respective UCL limits in both D- and Q-statistics charts, and are therefore assumed to describe common-cause variation. The model can be updated periodically by including new predicted samples already accepted (lying within the confidence limits). In this way variations such as seasonal changes can be incorporated in the model, making it more robust against false positive alarms.

4.3. Application phase

To demonstrate the sensitivity of this chemometric quality control of chromatographic data, a validated chromatogram was manipulated. This was done by adding a 0.1% area impurity peak hidden under the target peak in nine varied resolutions from 0 to 2 as illustrated in Fig. 8.

The nine simulated chromatograms were used to evaluate the methods ability to detect more or less hidden unexpected peaks. As indicated in the D-statistic chart (Fig. 9) none of the simulated chromatograms were detected, whereas in the Q-statistic chart (Fig. 9) chromatograms added impurity peaks with a resolution down to 1.5 was detected as faulty, falling outside the 3σ UCL.

It would be difficult or impossible to detect such an impurity peak visually or to identify it by peak integration using existing software. Generic peak detection algorithms commonly seek instants of rapid increase or decrease in signal intensity above a critical threshold. However, setting the threshold is a problem because too low a threshold generates a large number of meaningless peaks and too high a threshold might miss an actual one [2].

To determine chromatographic variables (retention time signals) responsible for the signal in the Q-statistic chart, a residual contribution plot is inspected in Fig. 10. The contribution plot allows us to diagnose the problem with the faulty chromatogram immediately. Clear indication of a new peak or a shoulder on the fronting target peak is given in Fig. 10. Apparently, this variability is not described by the principal components retained in the NOC model. Accordingly the added impurity with resolution 1.4 shows

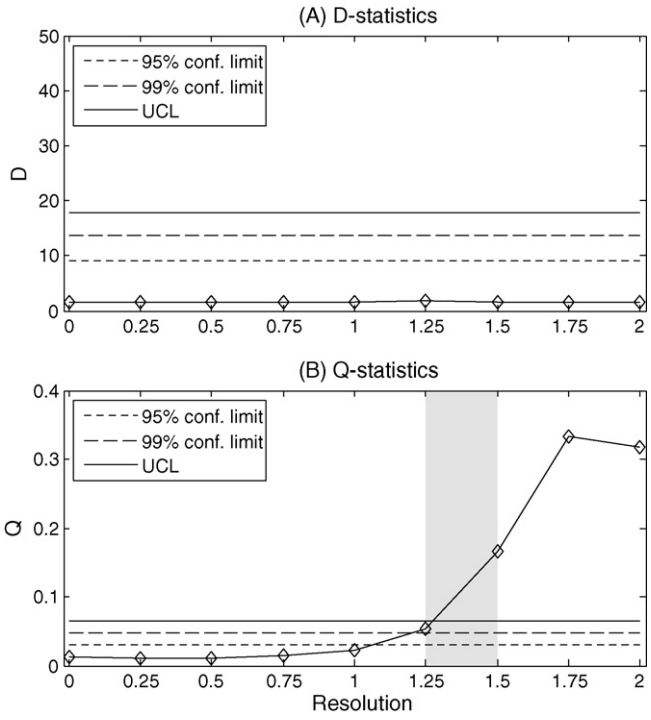

Fig. 9. Plot of D-statistics (A) and Q-statistics (B) of chromatograms added 0.1% area impurity with varying resolution (R_s 0–2). Critical area of detection in Q-statistics is marked.

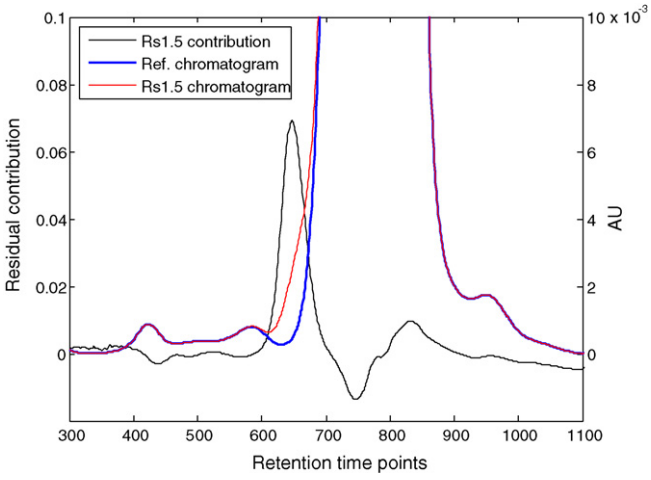

Fig. 10. Plot of the faulty R_s 1.5 residual contribution (black), plotted together with the reference (blue) and the faulty R_s 1.5 chromatogram (red) on the secondary *y*-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

up as an abnormal residual variability and a faulty signal in the Q-statistic chart.

5. Conclusions and perspectives

This study demonstrates that MSPC based on PCA applied on chromatographic purity analysis is a powerful tool for monitoring subtle changes in the chromatographic pattern. In addition it was illustrated how contribution plots provides clear diagnostics of faults at a glance. The chemometric quality control proved robust towards treating chromatographic artifacts such as baseline- and retention time drift. Applying this procedure for the detection of new peaks makes a fully automatic monitoring of complex chro-

matograms possible. Furthermore, if implemented and operating while the chromatographic purity analyses runs, this tool may considerably reduce time needed for subsequent assessment of peak integration. Thus, the chemometric quality control will increase throughput in chromatographic purity analysis and operate according to the process analytical technology (PAT) concept aiming for real-time release. The actual root cause of the alarm is not automatically given when applying chemometric quality control to HPLC purity analysis. Such an analysis would need incorporation of chemical and technical process knowledge or even more advanced analytical techniques e.g. coupled separation systems. Multivariate chromatographic patterns may well be increasingly important in the pharmaceutical industry. However, if the chemometric quality control described in this paper where to be integrated within the pharmaceutical industry, data management including smooth data accessibility will be a crucial requirement. Future work should be focused on incorporating the chemometric quality control in commercial software packages for chromatographic instruments or as part of a corporate database management system.

References

[1] International Conference on Harmonization (ICH), Guidance for Industry: Q3B(R2) Impurities in New Drug Products, 2006.
[2] B. Steffen, K.P. Müller, M. Komenda, R. Koppmann, A. Schaub, J. Chromatogr. A 1071 (2005) 239.
[3] R.L. Garnick, N.J. Solli, P.A. Papa, Anal. Chem. 60 (1988) 2546.
[4] Council of Europe, European Pharmacopoeia, 2007.
[5] United States Pharmacopoeial Convention Inc., US 26-NF 21, 2003.
[6] F.C. Sanchez, P.J. Lewi, D.L. Massart, Chemom. Intell. Lab. Syst. 25 (1994) 157.
[7] L. Zhu, R.G. Brereton, D.R. Thompson, P.L. Hopkins, R.E.A. Escott, Anal. Chim. Acta 584 (2007) 370.
[8] S. Ahuja, Adv. Drug Deliv. Rev. 59 (2007) 3.
[9] A.J. Lau, B.H. Seo, S.O. Woo, H.L. Koh, J. Chromatogr. A 1057 (2004) 141.
[10] H. Hotelling, J. Educ. Psychol. 24 (1933) 417.
[11] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.
[12] T. Kourti, Anal. Bioanal. Chem. 384 (2006) 1043.
[13] S. Kittiwachana, D.L.S. Ferreira, L.A. Fido, D.R. Thompson, R.E.A. Escott, R.G. Brereton, J. Chromatogr. A 1213 (2008) 130.
[14] T. Skov, R. Bro, Anal. Bioanal. Chem. 390 (2008) 281.
[15] R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde, M. van der Werf, BMC Genomics 7 (2006) 142.
[16] H.J. Ramaker, E.N.M. van Sprang, S.P. Gurden, J.A. Westerhuis, A.K. Smilde, J. Process Contr. 12 (2002) 569.
[17] F. Gan, G. Ruan, J. Mo, Chemom. Intell. Lab. Syst. 82 (2006) 59.
[18] P.H.C. Eilers, Anal. Chem. 76 (2003) 404.
[19] F. van den Berg, Baseline_spline: determines spline-based baseline by gradually eliminating points, unpublished work, 2008.
[20] F. van den Berg, G. Tomasi, N. Viereck, Warping: investigation of NMR pre-processing and correction, in: S.B. Engelsen, P.S. Belton, H.J. Jakobsen (Eds.), Magnetic Resonance in Food Science: The Multivariate Challenge, Royal Society of Chemistry, Cambridge, 2005, pp. 131–138.
[21] F. Savorani, G. Tomasi, S.B. Engelsen, J. Magn. Reson. 202 (2010) 190.
[22] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[23] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[24] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemom. 20 (2006) 484.
[25] M. Daszykowski, B. Walczak, J. Chromatogr. A 1176 (2007) 1.
[26] D. Bylund, R. Danielsson, K.E. Markides, J. Chromatogr. A 915 (2001) 43.
[27] A. Ferrer, Quality Eng. 19 (2007) 311.
[28] T. Kourti, J.F. MacGregor, Chemom. Intell. Lab. Syst. 28 (1995) 3.
[29] A. Nijhuis, S. de Jong, B.G.M. Vandeginste, Chemom. Intell. Lab. Syst. 38 (1997) 51.
[30] H. Hotelling, in: C. Eisenhart, M.W. Hastey, W.A. Wallis (Eds.), Techniques of Statistical Analysis, McGraw-Hill, New York, 1947, p. 113.
[31] J.E. Jackson, A user's guide to principal components, John Wiley and Sons, 1991.
[32] P. Nomikos, J.F. MacGregor, Technometrics 37 (1995) 41.
[33] S. Joe Qin, J. Chemom. 17 (2003) 480.
[34] J.E. Jackson, G.S. Mudholkar, Technometrics 21 (1979) 341.
[35] P. Miller, R.E. Swanson, C.E. Heckler, Int. J. Appl. Math. Comp. 8 (1998) 775.
[36] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Chemom. Intell. Lab. Syst. 51 (2000) 95.
[37] J.F. MacGregor, T. Kourti, Control Eng. Pract. 3 (1995) 403.
[38] P. Ralston, G. DePuy, J.H. Graham, ISA Trans. 40 (2001) 85.
[39] R.G. Harrison, P. Todd, S.R. Rudge, D.P. Petrides, Bioseparations Science and Engineering, Oxford University Press, New York, 2003.
[40] International Union of Pure Applied Chemistry (IUPAC), Pure Appl. Chem. 65 (1993) 819.