FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

USEFUL ~~NEW~~ TOOLS FOR ASSESSMENT OF CHROMATOGRAPHIC DATA

–A PHARMACEUTICAL CASE STUDY

PhD thesis by
Kristoffer Laursen · 2011

University of Copenhagen ▪ Faculty of Life Sciences
Department of Food Science ▪ Quality and Technology
Rolighedsvej 30 ▪ 1958 Frederiksberg C ▪ Denmark

Novo Nordisk A/S
DAPI MDEV ▪ Modeling and Optimization
Novo Allé ▪ 2880 Bagsværd ▪ Denmark

FACULTY OF LIFE SCIENCES

UNIVERSITY OF COPENHAGEN

Title:

Useful tools for assessment of chromatographic data
-a pharmaceutical case study

Supervisors:

Professor Rasmus Bro
Quality and Technology, Department of Food Science
Faculty of Life Sciences, University of Copenhagen, Denmark

Director Casper Leuenhagen
Modeling & Optimization, DAPI MDEV
Novo Nordisk A/S, Denmark

SVP cLEAN Coordinator Mads Thaysen
DAPI Coordination Group
Novo Nordisk A/S, Denmark

Opponents:

Associate Professor Jan H. Christensen
Department of Basic Sciences and Environment
Faculty of Life Sciences, University of Copenhagen, Denmark

Innovation Manager John Sørensen
Innovation Centre Nr. Vium
Arla Foods Amba, Denmark

Managing Consultant Eric van Sprang
TIPb - Toegepaste Industriële ProcesBeheersing
Amsterdam, Netherlands

PhD Thesis · 2011 © Kristoffer Laursen
Cover illustration by Magnus Dinesen
Printed by SL Grafik, Frederiksberg C, Denmark
ISBN: 987-87-7611-415-2

changing diabetes®

novo nordisk®

Preface

This PhD thesis is written to fulfill the requirements for obtaining a PhD degree at the University of Copenhagen, Faculty of Life Sciences. The presented work has been carried out at Modeling & Optimization, Diabetes API Manufacturing Development, Novo Nordisk A/S, and at Quality & Technology (Q&T), Department of Food Science, Faculty of Life Sciences, University of Copenhagen. The project has been sponsored by Novo Nordisk A/S, which is greatly appreciated. The project has been supervised by Professor Rasmus Bro from University of Copenhagen as well as Director Casper Leuenhagen and SVP cLEAN coordinator Mads Thaysen from Novo Nordisk A/S.

I am especially grateful to all my supervisors. The combination of your various professional competences and personal qualities has provided me the best possible supervision. You have all inspired me and supported me in different ways and different levels. It has been a true pleasure working with you. Thank you!

The presented papers involved several co-authors. Special gratitude is shown to Specialist Søren S. Frederiksen from Novo Nordisk A/S for sharing his deep chromatographic knowledge with me and helping me out with MATLAB issues, all of which supported the preparation of PAPER II. I also wish to thank Area Specialist Ulla Justesen from Novo Nordisk A/S for involving me in her project activities on LC-MS, which resulted in the preparation of PAPER III. Finally, I wish to thank PhD student Morten A. Rasmussen for his support and friendship, especially his statistical inputs which contributed to the preparation of PAPER I and PAPER III.

I am thankful to all my colleagues at Novo Nordisk A/S and Q&T for contributing to a pleasant, welcome and professionally fruitful working environment. Special thanks to Henrik Westborg and Niels V. Hartvig from Novo Nordisk A/S for inspiring discussions on multivariate statistics. I also wish to thank Thomas Skov for sharing his knowledge with me on chromatographic data and preprocessing possibilities.

The final appreciation goes to my family and friends for the sympathy and support. Special thanks to my wife Hanne, who enriched my life with a son half way through this PhD mission.

Kristoffer Laursen

Frederiksberg, February 2011

Abstract

The use of chromatographic analytical techniques for in-process monitoring of impurities is crucial for ensuring the purity of the final pharmaceutical product and thereby protecting the patient who ultimately receives it. Today it is industrial practice to assess chromatographic data by commercial chromatographic software combined with visual inspection of chromatograms and peak tables. Although simple, this partly manual method is quite laborious, extremely time consuming, seldom quantitative and prone to subjective decision-making. Subsequently, the levels of each impurity are monitored in separate control charts which make it difficult to detect if the relationship between impurities varies. Ultimately to guarantee that all impurities are resolved from the target compound and detected, it is desirable to add a further dimension to the chromatographic separation such as liquid chromatography-mass spectrometry (LC-MS). However, the relevant chemical information may not be easily accessible from the huge amount of data generated with LC-MS analysis, especially when the presence of unknown impurities is investigated.

The purpose of this PhD study has been to explore the potentials of new tools for improved assessment and monitoring of analytical chromatographic data from process samples in the pharmaceutical industry. This thesis demonstrates how newly developed methods and algorithms can ensure better utilization of available information in chromatographic data. The approach taken here, includes preprocessing of collected data in numerical software to generate 'clean' data; followed by multivariate statistical modeling that allows comprehensive control chart monitoring; and finally interpretable visualizations providing diagnostic information on deviating chromatographic data. Consequently, all these new and useful tools have been presented, explained and visualized on actual pharmaceutical analytical chromatographic data with more detailed information found in the attached scientific PAPER I-III.

In the first PAPER (PAPER I), a new comprehensive control (COCO) chart procedure is developed that considers both univariate statistics and multivariate statistics derived from PCA in a single plot that allows easy

visualization of the combined data from a univariate and multivariate point of view. The method is exemplified using integrated areas of analytical chromatographic peaks.

PAPER II proposes a powerful multivariate statistical process control (MSPC) approach based on principal component analysis (PCA) for monitoring subtle changes in the chromatographic profile. Clear diagnostic visualizations indicate subtle chromatographic deviations due to new impurities co-eluting with the target compound. The procedure supports the current practiced visual inspection of chromatograms by an automated and timely tool for continuous quality verification of chromatographic data in an objective and statistical reliable way.

In PAPER III, an MSPC tool based on PCA in conjunction with multiple testing is developed to adapt the nature of LC-MS data and applied to in-process LC-MS analysis of an industrial insulin intermediate. The tool detected low spike-levels (0.05%) of a structurally related compound co-eluting with the target compound and further provided clear diagnostics of the co-eluting compound. This tool makes a fully automatic monitoring of LC-MS data possible, where only relevant areas in the LC-MS data are highlighted for further interpretation.

In PAPER II and III, different chromatographic data preprocessing methods such as time alignment, baseline correction and scaling are applied to correct for non-relevant analytical variation, since it largely influences the outcome of the monitoring procedure.

In conclusion the research presented in this thesis has demonstrated the unique potentials of assessing chromatographic data using novel multivariate statistical tools. These tools utilize the available information contained in multiple measured chromatographic signals simultaneously in an objective (numerical) and statistically reliable way. The applications described in PAPER I-III may all serve as good alternatives or supplements to current procedures used in the pharmaceutical industry.

Resumé

Anvendelsen af kromatografiske analyseteknikker til procesovervågning af urenheder i lægemidler er af afgørende betydning for at sikre renheden af produktet og i sidste ende beskytte patienten. I industrien er det generel praksis at vurdere kromatografiske data ved hjælp af instrumentets indbyggede software kombineret med visuel vurdering af kromatogrammer og tabeller over de integrerede toppe. Denne enkle, delvis manuelle metode er meget tidskrævende, sjældent kvantitativ og er desuden påvirket af en subjektiv beslutningstagen. Niveauet af de enkelte urenheder overvåges ofte i separate kontrol kort, hvilket gør det vanskelig at opdage, hvis forholdet mellem urenheder varierer. For i sidste ende at sikre at alle urenheder er separeret fra hovedkomponenten og detekteret, er det ønskeligt at tilføje en ekstra dimension til den kromatografiske separation, såsom væskekromatografi-massespektrometri (LC-MS). Det er dog ikke altid let et ekstrahere relevant kemisk information fra den enorme mængde af data der genereres ved LC-MS analyse. Dette kan især være et problem når formålet er at undersøge tilstedeværelsen af ukendte urenheder.

Formålet med dette PhD-studium har været at udvikle forbedrede metoder til evaluering og overvågning af analytiske kromatografiske data fra procesprøver i den farmaceutiske industri. Denne afhandling viser, hvordan nyudviklede metoder og algoritmer kan sikre en bedre udnyttelse af den tilgængelige information i kromatografiske data. Fremgangsmåden indbefatter forbehandling af data for at generere "rene" data, efterfulgt af multivariat statistisk modellering, opsætning af kontrolkort der tillader en alsidig overvågning og endelig let fortolkelige visualiseringer, der giver diagnostisk information om afvigende kromatografiske data. Disse nye og nyttige metoder er blevet præsenteret, forklaret og visualiseret på faktiske farmaceutiske analytiske kromatografiske data. Flere detaljer kan findes i de vedlagte videnskabelige artikler (PAPER I-III).

I den første artikel (PAPER I), er en ny alsidig kontrolkort (COCO) procedure udviklet, som både håndterer univariat statistik og multivariat statistik vha. principal komponent analyse (PCA) i et enkelt plot. Dette COCO kontrolkort gør det nemt at visualisere data fra et kombineret

univariat og multivariat synspunkt. Metoden er eksemplificeret på integrerede arealer af analytiske kromatografiske toppe.

Den anden artikel (PAPER II) omhandler multivariat statistisk proces kontrol (MSPC) som er en fremgangsmåde baseret på PCA til overvågning af små ændringer i den kromatografiske profil. Diagnostisk visualisering indikerer små afvigelser i kromatogrammet på grund af nye urenheder der eluerer samtidigt med hoved komponenten. Artiklen beskriver hvordan den praktiserede visuelle inspektion af kromatogrammer kan understøttes med denne automatiserede og rettidige procedure til løbende kvalitets verifikation af kromatogrammer på en objektiv og statistisk pålidelig måde.

I den tredje artikel (PAPER III), er et MSPC værktøj baseret på PCA kombineret med multipel testning udviklet til LC-MS data, og anvendt til procesanalyse af et industrielt insulin mellemprodukt. Værktøjet er i stand til at detektere et lavt spike-niveau (0,05%) af et strukturelt beslægtet stof, der eluerer samtidigt med hovedkomponenten. Dette værktøj gør en fuldautomatisk overvågning af LC-MS data mulig, hvor kun relevante områder i data er fremhævet til yderligere fortolkning.

I den anden og tredje artikel (PAPER II–III), er der anvendt forskellige kromatografiske forbehandlingsmetoder såsom justering af tids-aksen, basislinie-korrektion og skalering med det formål at korrigere for irrelevant analytisk variation, da den i vid udstrækning påvirker resultatet af overvågnings-proceduren.

Resultaterne som er præsenteret i denne afhandling viser hvorledes kromatografiske data kan vurderes ved hjælp af nye multivariate statistiske redskaber. Disse værktøjer udnytter informationen i de multiple kromatografiske signaler på en objektiv, datadrevet og statistisk pålidelig måde. Metoderne, beskrevet i de tre artikler er alle udviklet som alternativer eller supplementer til de nuværende metoder, der anvendes i den farmaceutiske industri.

List of publications

The following scientific papers are discussed in this thesis. They are referred to in the text by their Roman numerals I-III.

PAPER I

Laursen K., Rasmussen M.A., Bro R.

Comprehensive control charting applied to chromatography

Chemometrics and Intelligent Laboratory Systems, 107 (2011) 215-225

PAPER II

Laursen K., Frederiksen S.S., Leuenhagen C., Bro R.

Chemometric quality control of chromatographic purity

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

PAPER III

Laursen K., Justesen U., Rasmussen M.A.

Enhanced monitoring of biopharmaceutical product purity using liquid chromatography - mass spectrometry

Journal of Chromatography A, 1218 (2011) 4340-4348

List of abbreviations

API	Active Pharmaceutical Ingredient
AU	Absorbance Unit
BPC	Base Peak Chromatogram
CL	Control Limit
COCO	COmprehensive COntrol
coshift	COrrelation-SHIFTing
COW	Correlation Optimized Warping
CUSUM	CUmulative SUM
DAD	Diode Array Detection
EIC	Extracted Ion Chromatogram
ESI	ElectroSpray Ionization
EWMA	Exponentially Weighted Moving Average
HPLC	High Performance Liquid Chromatography
ICH	International Conference on Harmonization
*i*coshift	Interval-COrrelation-SHIFTing
ITA	Initial Training Application
LC	Liquid Chromatography
LCL	Lower Control Limit
LC-MS	Liquid Chromatography Mass Spectrometry
MS	Mass Spectrometry
MSPC	Multivariate Statistical Process Control
NMR	Nuclear Magnetic Resonance
NOC	Normal Operation Condition
PARAFAC	PARAllel FACtor analysis
PAT	Process Analytical Technology
PC	Principal Component
PCA	Principal Component Analysis

PLS	Partial Least Squares
QbD	Quality by Design
RMSE	Root Mean Square Error
RMSEC	Root Mean Square Error of Calibration
RMSECV	Root Mean Square Error of Cross Validation
RMSEP	Root Mean Square Error of Prediction
SPC	Statistical Process Control
sqrt	SQuare RooT
std	STandard Deviation
SVD	Singular Value Decomposition
TIC	Total Ion Chromatogram
TOF	Time Of Flight
UCL	Upper Control Limit
UV	Ultra Violet

Table of contents

1 Introduction

This PhD project was driven by a need and desire to develop alternative solutions providing simpler yet more comprehensive monitoring capabilities of analytical chromatographic data in the pharmaceutical industry. The use of chromatographic analytical techniques for in-process monitoring of impurities, is crucial for ensuring the purity of the final pharmaceutical product, and thereby protecting the patient who ultimately receives it. Today, it is industrial practice to monitor each impurity of interest with a separate control chart, which indicates the range of acceptable variation in concentration of the impurity. However, for in-process analysis, several impurities may be of interest and this will necessitate that the practitioner inspects a large number of control charts. Moreover, when special events occur in a process they affect not only the magnitude of the impurities, but also their relationship to each other. These events are often difficult to detect by charting one impurity at a time because the correlations between the impurities are not directly reflected in the individual control charts [PAPER I].

Often analytical chemists and laboratory technicians are limited to the integration systems available in the commercial chromatographic software. This software often suffer from low reliability towards identifying unknown peaks when these have low signal to noise ratio and are overlapping with other peaks. Thus, it is common practice to assess the results of peak integration by visual inspection of the chromatogram. Visual inspection of chromatograms have been used for decades [1], and is a valid procedure for identification of protein samples recognized by the regulatory authorities [2,3]. Although simple, this partly manual method is quite laborious, extremely time consuming, seldom quantitative and prone to subjective decision-making [PAPER II].

Although high-performance liquid chromatography (HPLC) is the most widespread analytical tool for in-process purity testing, it is recognized that HPLC can not guarantee that all impurities are resolved from the target compound (usually present in excess compared to any impurity). It is therefore desirable to add a further dimension to the chromatographic separation to increase confidence that all impurities are detected. Coupling

mass spectrometry (MS) to liquid chromatography (LC-MS) adds a more selective dimension to the chromatographic separation. However, the relevant chemical information may not be directly accessible from the huge amount of data generated with LC-MS analysis. That is particularly when the presence of unknown impurities is investigated, which can be considered as a case of needle-in-the-haystack expedition, due to the nature of LC-MS data [PAPER III].

It would be of major benefit for the pharmaceutical industry if these challenges could be handled by new and useful tools, improving assessment of chromatographic data and providing more comprehensive monitoring capabilities. In the following some motivations for this PhD project are described, and finally the aim and outline of this thesis is given.

1.1 Scientific motivations

Pharmaceutical process and product monitoring demands an array of in-process analyses, which consequently generate a huge amount of data, containing hundreds or even thousands of variables. Despite significant benefits may be gained from such analytical data; it is generally not a trivial task to extract relevant information and knowledge from these data. Thanks to the development of computer power and multivariate statistical data analysis, spectacular progress has been achieved in comprehensively extracting relevant information from analytical data. The pharmaceutical industry can benefit from the wealth of knowledge accumulated and published over the years within the field of multivariate statistical data analysis. These methods have been successfully applied in other industries and research areas. If the multivariate statistical analysis philosophy is adapted by the pharmaceutical industry and further developed into industrially reliable on-line monitoring schemes, it can lead to more powerful and applicable methods, which can become useful to a broader range of users.

1.2 Industrial motivations

Today many people depend upon the quality of pharmaceutical products for their everyday health care. Pharmaceutical products are expected to be safe and efficient whenever needed – day after day, year after year. If the pharmaceutical product quality fails, the consequences can be catastrophic

leading to annoyance, inconvenience and even more severe effects on the customer. It takes a long time for a company to build up a reputation of reliability, and only a short time to be branded as "unreliable". Therefore, continual assessments of product quality are a critical necessity in the pharmaceutical industry. In light of the recent quality by design (QbD) initiative by the U.S. Food and Drug Administration (FDA) [4], increasing attention has been drawn to the application of the QbD principles [5,6] for impurity investigation and control, emphasizing process understanding based on sound science and risk management [7-9]. Under the new QbD paradigm, impurities should not only be tested for in the end product, but rather be proactively controlled by design throughout the manufacturing process. This of course, requires powerful analytical techniques and comprehensive extraction of relevant information from the analytical data that would govern early warnings of deviating product quality. Early warnings may lead to timely corrections and consequently a minimization of the number of rejected batches, product rework and lengthy failure investigations. Moreover, improved monitoring of product purity will lead to more effective and less complicated risk management procedures e.g. during changes and optimizations of processes.

For a pharmaceutical company, all of these benefits will ensure license to operate and could furthermore result in major savings and additional funding for research and development of new and better products for the benefit of the patient.

1.3 Aim of thesis

This thesis focuses on solutions for more comprehensive monitoring capabilities of analytical chromatographic data in the pharmaceutical industry, which simply ensure better utilization of available information in chromatographic data. Therefore, the aim of this thesis is to develop methods and algorithms that improve assessment and monitoring of chromatographic data obtained for purity analysis in the pharmaceutical industry. The intention is to automate and optimize the many aspects present when setting up an industrial reliable monitoring scheme as illustrated in Figure 1. This includes: Collection of data from commercial chromatographic instruments to numerical software such as MATLAB (MathWorks); application of necessary preprocessing steps to generate ‘clean’ data for subsequent modeling; multivariate statistical modeling

representing the critical chromatographic data; comprehensive control chart monitoring; and interpretable visualizations providing diagnostic information on deviating data.

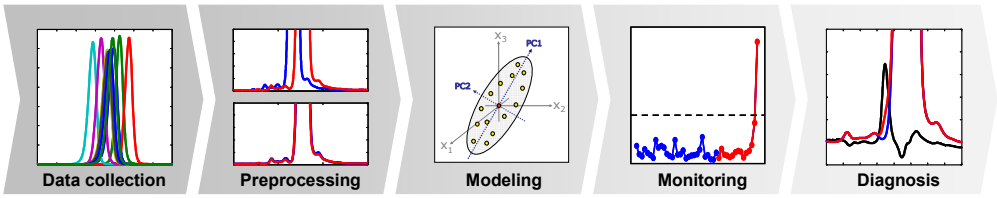

Figure 1. Aim of thesis; from collection of chromatographic data, preprocessing for subsequent multivariate statistical modeling, followed by control chart monitoring, and finally diagnostic information.

1.4 Thesis outline

The thesis consists of an introductory part followed by three scientific papers (PAPER I, II and III). The introductory part serves to introduce the reader to pharmaceutical product purity; the methods used in the study as well as the major results, and are organized as follows:

1.4.1 Chapter 2 Pharmaceutical product quality

Chapter 2 describes the importance of assuring pharmaceutical product quality by monitoring impurities. Analytical chromatographic methods commonly used for purity analysis in the pharmaceutical industry are introduced, and state-of-the-art monitoring systems used to asses the qualified status of the product during production are reviewed. Finally, this chapter gives an overview of the multivariate monitoring scheme used throughout this thesis.

1.4.2 Chapter 3 Chromatographic data

This chapter gives a brief historic perspective on the development of chromatography, and on the instrumental hyphenation properties with mass spectrometry. The data structure and dimensionality of HPLC data (univariate UV detection) and LC-MS data (multivariate mass detection) will be discussed, and some aspects of chromatographic peak resolution and peak purity will be touched upon.

1.4.3 Chapter 4 Preprocessing

Chapter 4 describes how various preprocessing methods can prepare the raw chromatographic data for subsequent multivariate statistical

monitoring. The selected preprocessing methods described here all found their usefulness in the papers included in the thesis. Among those are; baseline correction, peak alignment, and scaling methods.

1.4.4 Chapter 5 Multivariate statistical monitoring

This chapter includes different aspects of multivariate statistical monitoring based on principal component analysis (PCA). The usefulness of monitoring chromatographic data in a multivariate statistical way will be discussed and examples will be given from the papers included in this thesis. Since the theory behind multivariate statistical monitoring originates from statistical process control (SPC) methodology, a brief introduction is given to the concepts of SPC and the link to multivariate SPC (MSPC).

1.4.5 Chapter 6 Conclusions and perspectives

Finally, Chapter 6 summarizes the major findings of new and useful tools which improve assessment of chromatographic data and provide more comprehensive monitoring capabilities in the pharmaceutical industry. The topics where additional work and focus is needed will be discussed.

2 Pharmaceutical product quality

Two branches exist in pharmaceutical production: the manufacture of active pharmaceutical ingredients (APIs), also known as drug substances, and the manufacture of drug products. With drug substance manufacturing, the active ingredient is synthesized during the course of many individual chemical reactions or process steps. Subsequently drug product manufacturing involves carrying out a formulation of the drug substance which delivers the drug substance in a stable, non-toxic and acceptable form, ensuring its bioavailability and therapeutic activity. This thesis will focus on the manufacture of the drug substance, as this branch has been the foundation for the scientific work carried out during this PhD project.

Safety and efficacy of pharmaceutical products are two fundamental quality issues of importance in pharmacotherapy (treatment of diseases through the use of pharmaceutical products). The safety of a pharmaceutical product is dependent not only on the toxicological properties of the active drug substance itself, but also, for example, on the impurities that it contains. Additionally, these impurities could potentially compromise the efficacy of the active drug substance [10]. Thus, the analytical activities concerning impurities in pharmaceutical products are among the most important issues in modern pharmaceutical analysis [11].

An impurity in a drug substance as defined by the International Conference on Harmonization (ICH) guideline document Q3A [12] as: *"any component of the drug substance that is not the chemical entity defined as the drug substance"*. Impurities in drug substances may originate from various sources and phases of the process. The origin of impurities will not be described further here, but several reviews offer insights into these matters [13-15]. Regulatory agencies also explicitly regulate the control criteria for these impurities in drug substances by providing guidance for the pharmaceutical industry. These are not discussed here but are outlined in the ICH Q3A guideline document [12]. The analytical testing for and evaluation of impurities are important requirements. This, of course, requires suitable and powerful analytical methods; these are briefly described in the following subsection. Finally, in order to maintain the qualified status of the product during production, the known impurities have to be monitored, and the unlikely

presence of new unknown impurities should preferably be detected as early as possible. Such impurity monitoring systems are discussed in the final subsection of this chapter.

2.1 Analytical methods for purity testing

As stated previously effective testing and monitoring of impurities is crucial for the pharmaceutical industry. This, of course, requires suitable and powerful analytical methods. Analytical testing of impurities in pharmaceutical products is also an important regulatory issue. The validation of analytical procedures, i.e., the proof of its suitability for the intended purpose, is an important part of the registration application for a new pharmaceutical product. The ICH has harmonized these requirements in the Q2(R1) guideline document [16].

Since impurities are usually present in relatively small quantities compared to the drug substance, an analytical technique capable of separating a mixture containing highly varied concentrations of analytes with sensitive and specific detection is required. Today, high-performance liquid chromatography (HPLC) with UV detection is the most commonly used analytical technique for purity testing of in-process intermediates and drug substances. HPLC has been the most important analytical method for determination of impurities in pharmaceutical products for over two decades [11]. However, it is recognized that HPLC can not guarantee that all impurities are resolved from the target compound usually present in excess compared to any impurity. It is therefore desirable to evaluate one or more complementary analytical methods to increase confidence that all impurities are detected and identified. The addition of further dimensions to chromatographic separations by hyphenated techniques offers unique opportunities for so-called peak-purity examination of the target compound. HPLC with diode array detection (HPLC-DAD) is a commonly used method to conduct peak-purity examination. However, many impurities are structurally related to the drug substance, and their structure may contain very similar chromophores, making purity assessment based solely on HPLC-DAD data difficult and unreliable. Coupling mass spectrometry (MS) to liquid chromatography (LC-MS) adds a more selective dimension to the chromatographic separation. Since MS separates compounds by their respective mass-to-charge ratios (m/z), any difference in the m/z values between the impurities and the active drug substance will allow an

unambiguous detection regardless of similarities in their UV spectra. Therefore an impurity co-eluting with the target peak will be separated in MS as long as their *m*/*z* values are different and ionization of the impurity is not suppressed by the target compound. LC-MS may also facilitate correct assignment of new peaks arising at the same retention time as known ones, which potentially are wrongly assigned with UV detection. Furthermore identification is improved by the use of LC-MS, as molecular masses are assigned to impurity peaks. In this way verification can also be provided as to whether impurities are really 'new' or whether they were already present in previous batches in lower amounts. This might help in the toxicological evaluation when taking, for example, safety factors into account.

2.2 Current monitoring systems

The requirement to show process and batch consistency demands an array of in-process analysis, which consequently generates a huge amount of data containing hundreds or even thousands of variables. These routinely measured data are automatically recorded in historical databases for the purposes of product monitoring, process control, and potentially process improvement/optimization.

For example, during the production of each batch, process operators and quality-control departments normally assess these analytical data to ensure the product quality and take appropriate corrective actions when needed. However, it is generally not a trivial task to assess these analytical data and utilize all the available information. Therefore, many pharmaceutical processes face a well known problem, i.e., 'data rich but information poor', despite that significant potential benefits may be gained from these data

Analytical data, if based on HPLC, usually consists of integrated areas of a number of well known peaks (the target compound and related impurity compounds). Most commonly batch-to-batch variation is analyzed on a less frequent basis (weekly, monthly, quarterly or once a year) to asses the long-term quality and stability of the product. Here, the concentration of each compound of interest is monitored with a separate control chart, which is a simple plot of the compound concentration vs. time, sample or batch. However, control chart monitoring of an in-process analysis containing several impurity compounds will force the practitioner to inspect a large number of control charts, and the risk of making mistakes is larger when several control charts are to be checked [17]. When special events occur in a

process they affect not only the magnitude of the compounds but also their relationship to each other. These events are often difficult to detect by charting one compound at a time because the correlations between the compounds is not directly affected in the individual control charts [PAPER I]. Another problematic issue is the use of generic peak detection algorithms which 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 HPLC chromatogram. As mentioned previously visual inspection of chromatograms is a valid procedure for identification of protein samples recognized by the regulatory authorities [2,3]. Although simple, this partly manually method is quite laborious, extremely time consuming, seldom quantitative and prone to subjective decision-making probably causing additional errors [PAPER II]. As for HPLC data, generic peak detection algorithms used for LC-MS data may also suffer from inconsistent reliability and thus manual interpretation is often necessary. However, manual interpretation of LC-MS data is extremely tedious; particularly in reference to applications where pharmaceutical product purity is monitored and unknown peaks are to be registered if present [PAPER III].

Obviously there are needs for more automatic and timely tools that can monitor these chromatographic data objectively, quantitatively, and in a statistically reliable way. Furthermore, these tools should automate the less frequent review of batch-to-batch variation and turn it into a continuous review. These needs are strongly supported by the increased focus on process analytical technology (PAT) [18] and quality by design (QbD) [4], which aims for enhanced process understanding that improves process control moving towards continuous quality verification and real-time release of an end product.

2.3 Multivariate monitoring scheme

In standard statistical process control (SPC) as well as multivariate SPC (MSPC) terminology the monitoring scheme is carried out in two distinct phases, Phase I and Phase II. In Phase I, a statistical model is constructed from a historical data set, which is assumed to be in control. In Phase II, the future observation is checked to see whether it fits well in the model. An extension to this standard monitoring scheme was proposed by H.J.

Ramaker et al. [19]. They carry out MSPC in three phases: The Initial, Training and Application phase, also referred to as ITA trajectory. Here, the training and application phase refer to respectively Phase I and Phase II from the standard terminology. In this study the multivariate statistical monitoring scheme of chromatographic data follows a modified version of the ITA trajectory as illustrated in Figure 2.

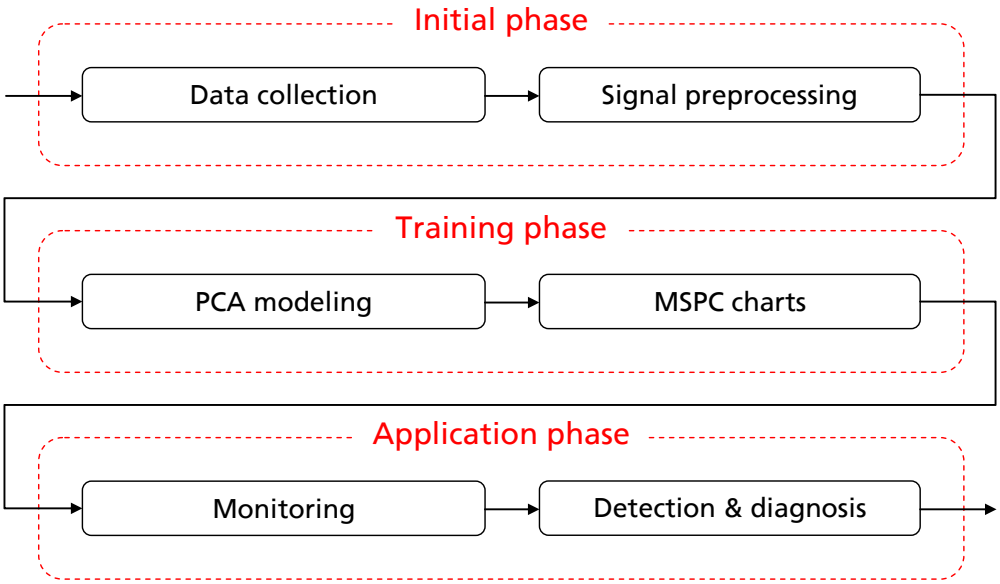

Figure 2. The three phases according to the ITA trajectory (Initial, Training and Application phase).

The three phases according to the ITA trajectory are described briefly in the following subsections.

2.3.1 Initial phase

In the initial phase, appropriate and representative historical chromatographic data are collected and preprocessed (described in Chapter 4). This historical dataset is one in which the process has been running consistently, under normal operation conditions (NOC), and only acceptable high quality products have been obtained. Normally, data are spread in various systems and are not always accessible in an easy manner. For instance the raw chromatographic signals most often have to be collected directly from the chromatographic instruments or a dedicated chromatographic database system. Quality measurements of the product,

including integrated peak areas, are usually stored in LIMS (Laboratory Information Management System). Therefore, the initial phase may often be a time consuming step if not automated.

2.3.2 Training phase

In the training phase a PCA model based on extracted and prepared NOC data is developed (describing common cause variation) and MSPC charts are constructed. Since this NOC model serves as a reference distribution and exclusively determines whether a new sample is similar or deviates significantly from the NOC samples, the monitoring performance depends very much upon adequacy and representativity of these NOC chromatograms. If e.g. a faulty chromatogram is included in the NOC model, the total amount of chromatographic variation increases, and the reference distribution now consists of non-representative NOC samples. Consequently, the model becomes less capable of detecting differences in variation between the NOC chromatograms and a new faulty chromatogram. Therefore, validation is an essential part of model development (described in subsection 5.4) to avoid false correlations, and to ensure that the estimated model reflects only NOC.

The number of samples needed to construct an adequate NOC model depends on the application. The effect of the size of the training set on the false alarm rate in statistical process monitoring have been investigated by Ramaker et al. [20].

2.3.3 Application phase

Finally, in the application phase new independent chromatographic data are prepared, fitted to the NOC model, and monitored using the control charts developed in the training phase. Deviating samples are diagnosed using contribution plots to determine causes of the deviating behavior. However, contribution plots do not automatically reveal the actual cause of the fault. Therefore, incorporation of chemical and technical process knowledge may be necessary to diagnose the problem and discover the root causes of the fault [21]. The NOC model can be updated periodically by including new samples already accepted by the NOC model (lying within the control limits) and with acceptable high product quality. In this way variations du to e.g. seasonal changes or different raw material suppliers can be incorporated in the NOC model, making it more robust against false positive alarms. However, if a consistent fault is detected and this fault is caused by

e.g. a permanent process change or a new raw material quality, the NOC model should be recalculated based on new NOC samples to reflect the present process conditions.

3 Chromatography

Testing, monitoring and evaluation of impurity profiles in pharmaceutical products are important regulatory requirements which, of course, require suitable and powerful analytical methods. Although HPLC is the most widespread analytical tool for purity testing, it is recognized that HPLC can not guarantee that all impurities are resolved from the target compound usually present in excess compared to any impurity. It is therefore desirable to add a further dimension to the chromatographic separation to increase confidence that all impurities are detected. Here liquid chromatography coupled to mass spectrometry (LC-MS) is a powerful and widely used analytical technique in the characterization and identification of impurities in pharmaceutical products.

The aim of this chapter is not to present chromatographic or mass spectrometric basic theory and instrumental setup or deal with how to optimize the chromatographic and mass spectral conditions for proper resolution and detection. This can be found in more dedicated textbooks [22-26]. The focus will be on chromatographic data representation and how to take full advantage of the available information hidden in the data structure and dimensionality of HPLC data (univariate UV detection) and LC-MS data (multivariate mass detection). Additionally, some aspects of peak co-elution and peak purity will be touched upon. To begin with a brief historic perspective on the development of HPLC and LC-MS is given.

3.1 Brief chromatographic history

The Russian botanist Mikhail Tswett is generally referred to as the father of chromatography. His work, originally presented in 1903 and then published in 1906 [27], described the separation of plant pigments by column liquid chromatography. Tswett defined the term chromatography, which originates from the two Greek words, chroma (color) and graphein (to write) [28]. Initially, not much attention was given to chromatography but, after a few decades, Tswett's discovery was re-considered by a few scientists and various modalities of chromatography emerged. Still, though the great discovery of Tswett, was not widely recognized. In 1941 Martin and Synge [29] published their Nobel Prize-winning article in which they introduced

liquid-liquid (or partition) chromatography and the accompanying theory that became known as plate theory. Later, Alm [30] reported the method of gradient elution in 1952. After these early developments, applications of liquid chromatography appeared more rapidly between 1960 and 1970 when high performance liquid chromatography (HPLC) was developed as an analytical tool in addition to gas chromatography [31]. Around 1973, packing technologies and development of reversed phase silica gel led to the first reversed phase HPLC columns [32]. Since then, several advancements have been developed for HPLC. Today reversed phase HPLC is a powerful tool for the modern laboratory and has played a key role as an analytical method in the development and control of pharmaceuticals.

For decades, the liquid chromatograph has been a working horse in the separation of compounds. At the same time the mass spectrometer has been an important and sensitive tool for structure elucidation. By hyphenating the two techniques, a very powerful instrumental set-up is achieved. Liquid chromatography-mass spectrometry (LC-MS) is an analytical technique that couples high resolution chromatographic separation with sensitive and specific mass spectrometric detection.

One of the first attempts at LC-MS was reported in 1968 by Talroze et al. [33], using a capillary inlet interface. In the 1980s several other type of interfaces was suggested, including, e.g., thermospray [34] and fast atom bombardment [35]. The breakthrough for LC-MS was, however, the development in the 1990s of two techniques for atmospheric pressure ionisation: the electrospray ionisation (ESI) [36] and the atmospheric pressure chemical ionization [37]. With ESI it is possible to obtain multiply charged ions for large molecules [38], e.g. proteins and carbohydrates. Thereby the detection of high molecular weight compounds is facilitated for instruments with limited m/z range. The technique is still developing, particularly in the mass spectrometry area, with vastly improved sensitivity and resolution. Today LC-MS is probably the most powerful technique available for pharmaceutical analysis, and the most common mass analyzers are those used in quadrupole, time of flight (TOF) and ion trap instruments [39].

3.2 Chromatographic data structure and dimensionality

In the case of HPLC, the separation and subsequent detection of compounds in a sample delivers a chromatogram. A chromatogram is a graphical

representation of all peaks eluting from the column superimposed on the baseline. The areas and heights of the peaks usually increase linearly with the amount of injected component [26]. Typical purity analysis in industrial processes deals with a manageable amount of compounds at relatively high concentrations. The original data obtained from the instrument can be transformed by integration of the peaks and the integrated data of selected peaks can then be used for subsequent data analysis. This can easily be handled automatically with available software packages suitable for routine analysis of chromatograms [40]. This is illustrated in Figure 3 where selected peaks of interest are marked and integrated in a HPLC chromatogram obtained for purity analysis of a biopharmaceutical in-process sample.

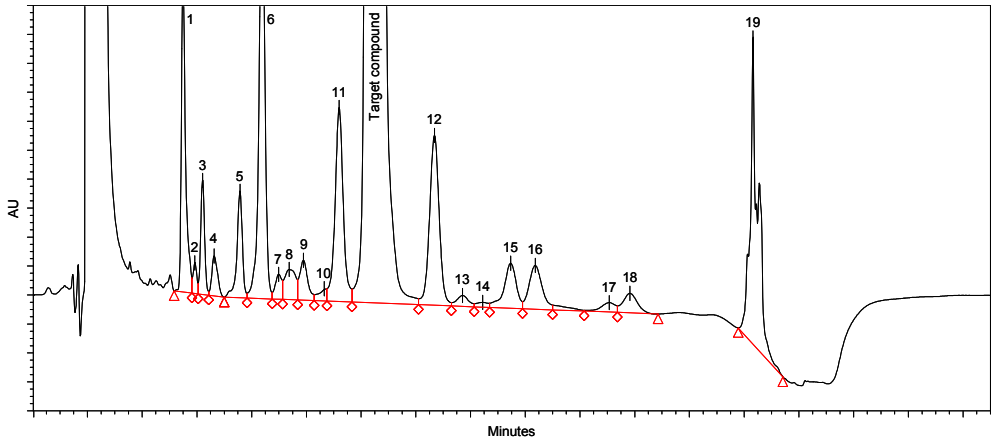

Figure 3. Analytical HPLC chromatogram of a biopharmaceutical in-process sample. Selected peaks are marked and integrated automatically [PAPER I].

However, generic peak detection algorithms often suffer from low reliability towards smaller peaks with low signal to noise ratio and overlapping peaks of different shapes. Hence, this data reduction can lead to a loss of information since the quality of the data relies on peak detection and on how the peaks are selected and integrated [41]. In other words, any error in the measurement of peak size will produce a subsequent error in the reported result. Therefore, it is common practice to assess the results of peak integration by visual inspection of the chromatogram. Although simple, this partly manual method is quite laborious, extremely time consuming, seldom quantitative and prone to subjective decision-making probably causing additional errors. As yet another alternative to automated peak detection and the laborious manual inspection, whole chromatographic profiles

collected from the instrument can be used mathematically, without first integrating a set of selected peaks. The result is that not only peak size is included but also its shape (peak overlap and peak shoulders). However, this requires uniform representation of the chromatographic signals in matrix form. As any other instrumental signal, the chromatographic profile contains three major components: the analytical relevant signal; the background or baseline; the noise. These are illustrated in Figure 4.

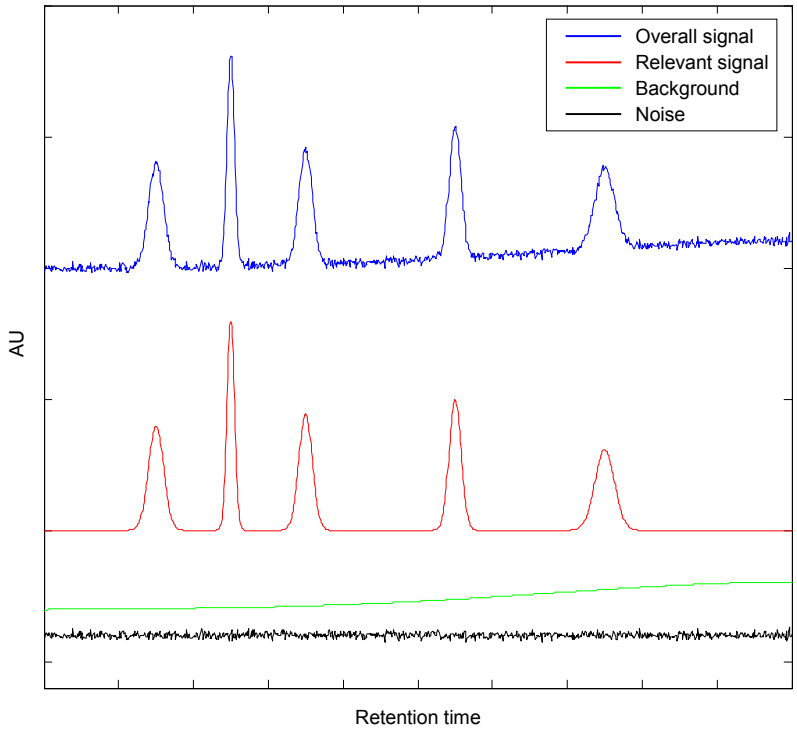

Figure 4. Components of the chromatographic analytical signal: (a) overall signal; (b) relevant signal; (c) background; and, (d) noise (visualization inspired by Daszykowski and Walczak (2006) [42]).

On top of these three different types of variation, there are also additional problems with chromatographic data. For example, the retention time of specific peaks can vary slightly from run to run for various reasons. Retention time shifts are problematic since they severely obscure comparison of chromatographic profiles. When whole chromatographic profiles are compared, these non-relevant components of Figure 4, and the shifting of chromatographic peaks need to be handled by the data analysis approach. In many cases, it is possible that the unwanted variation can be

corrected for prior to multivariate statistical monitoring. This can be done using suitable preprocessing as explained in Chapter 4.

Both peak tables and chromatographic profiles are considered two-way data and can be organized as an $M \times N$ data matrix, with M samples and N peak areas or elution time points (also referred to as retention time points). This matrix structure can readily be used as input for two-way multivariate statistical monitoring, described in Chapter 5.

In the case of LC-MS, the separation and subsequent mass spectral detection of compounds in a sample delivers a data matrix characterized by the intensity as a function of retention time and m/z (Figure 5A). Analysis by LC-MS can generate huge amounts of data, especially when the MS is operated in the full-scan mode over large regions in m/z.

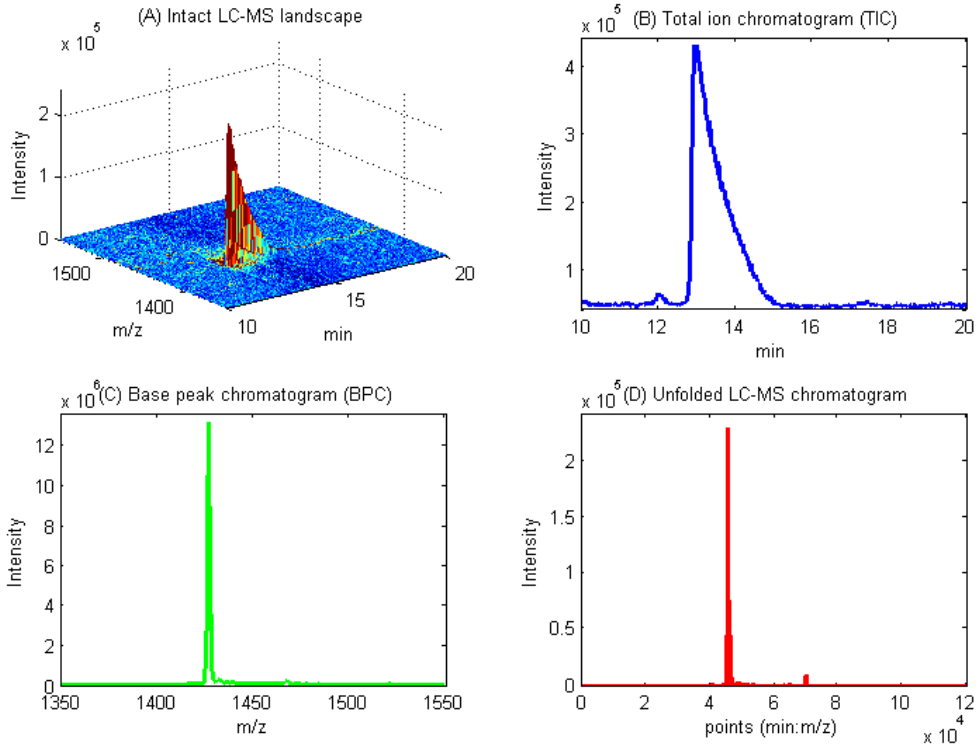

Figure 5. Different LC-MS data structure presentations: (A) intact LC-MS landscape; (B) total ion chromatogram (TIC); (C) base peak chromatogram (BPC); (D) unfolded LC-MS chromatogram (modified from PAPER III).

As shown in Figure 5, different data structures can be extracted from a single LC-MS sample:

(A) Intact landscape holding all available information

(B) Elution time profile (summed MS dimension – denoted total ion chromatogram (TIC))

(C) Mass spectral profile (summed LC dimension – denoted base peak chromatogram (BPC))

(D) Unfolded LC-MS chromatogram (the sample matrix is rearranged into a vector by concatenating the rows or column; here the m/z rows are concatenated)

As for HPLC data, automatic peak detection algorithms for LC-MS data may also suffer from low reliability and thus manual interpretation is often necessary. However, manual interpretation of LC-MS data is extremely tedious, particularly in reference to applications where pharmaceutical product purity is monitored and unknown peaks are to be registered if

present. If LC-MS data are to be compared by two-way multivariate statistical monitoring, then one dimension must be reduced either by summing or unfolding as illustrated in Figure 5. By summing, the amount of data points is simply reduced, whereas with unfolding the amount of data points is kept intact. Alternatively the intact LC-MS landscapes can be compared by advanced so-called multiway statistical methods (also referred to as factor models) such as PARAFAC [43]. These methods give new possibilities with regard to the information that can be extracted, but are not as widespread and user-friendly as two-way methods, due to their more sophisticated nature.

3.3 Peak detection and integration errors

The detection of peaks in a chromatogram is crucial for both qualitative and quantitative analyses, for the amount of information increases as more peaks are detected. However, peak overlap and baseline noise make the detection of peaks rather problematical. For instance a false peak may be detected when baseline noise might be taken for a minor peak, or a peak may be lost when the occurrence of overlapping is not recognized. Most routinely used detection methods do not employ any assumption for peak shapes or baseline noise. Most often the derivatives of the signal are analyzed and a peak is detected when a threshold is exceeded. All the information these peak detection methods use is that a peak is a signal that goes up and comes down [44]. Quantitative determination of the individual compounds can simply be done by integration of the peak area. For a correct area determination, the location of the baseline, the values of peak height and peak width must be measured with high precision. Baseline noise, drifting baseline, peak tailing or fronting, and peak overlap, all influence the accuracy and precision of the measurements made on chromatographic peaks. For high-purity pharmaceutical products, the target compound is present in excess compared to a potential impurity. Specifically, small peak size ratios from about 5% to less than 0.5% of the target peak commonly occur in the determination of impurities in pharmaceutical products [45]. When such a small impurity peak elute near the much larger target peak, situations will occur in which the small peak cannot be integrated as a separate peak because a valley no longer appears between the peaks. In such situations, careful examination of the baseline is necessary to determine the correct location for the integration start-stop positions. However, the use of rather simple and inappropriate integration methods (often implemented in

commercial chromatographic software) may result in underestimation or overestimation of peak areas. The integration errors are likely to occur due to asymmetry of one or both peaks (e.g. tailing) [45]. Mathematical peak models can be used to resolve the overlapping peaks into pure peak profiles (also referred to as peak deconvolution). Several peak fitting algorithms and procedures are available, but they are outside the scope of this thesis.

3.4 Peak resolution

The resolution expresses the extent of separation between the components in a sample, and is a useful measure of the separation properties of the column for a particular sample. The higher the resolution of peaks in the chromatogram, the better separation of the components the column provides. The separation ability of a column is characterized by the plate number, which determines the peak width relative to the retention time. A simplified method to calculate the resolution of a chromatogram is to use the plate model [46]. 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 chromatogram 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 [47]. The parameter σ is calculated as w_b divided by four. This is illustrated in Figure 6.

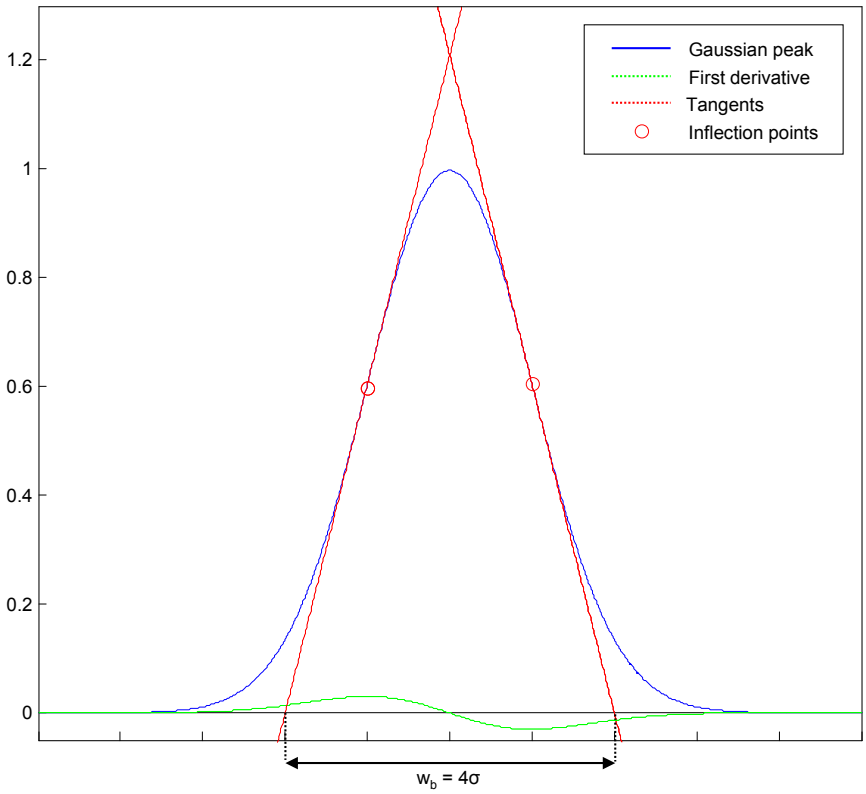

Figure 6. Width of a Gaussian peak, as a function of the standard deviation of the peak (modified from Ettre (1993) [47].

To define to what extent an impurity is hidden under the target peak; the peak resolution (R_s) is used. R_s expresses the efficiency of separation of two peaks in terms of their average peak width at base [47]:

$$R_s = 2\frac{(t_{R2} - t_{R1})}{(w_{b1} + w_{b2})} \quad (1)$$

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 [47]:

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

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 [47]:

$$R_s \approx \frac{(t_{R2} - t_{R1})}{w_{b2}} \tag{3}$$

In Figure 7 different degrees of chromatographic resolution is illustrated. Impurity peaks at 0.1% of the target peak area were simulated based on the assumptions in Equation 6. The impurity peaks were generated as pure Gaussian peaks using σ estimated from the target peak. Impurities were simulated with varied resolutions (R_s) from 1 to 2 (eluting after the target peak) and added the target peak chromatogram. In the upper plots in Figure 7 (A1 to A3) a symmetric target peak is added a 0.1% impurity peak with resolutions from 1 to 2, whereas in the lower plots (B1 to B3) an asymmetric (tailing) target peak is added a 0.1% impurity peak with resolutions from 1 to 2.

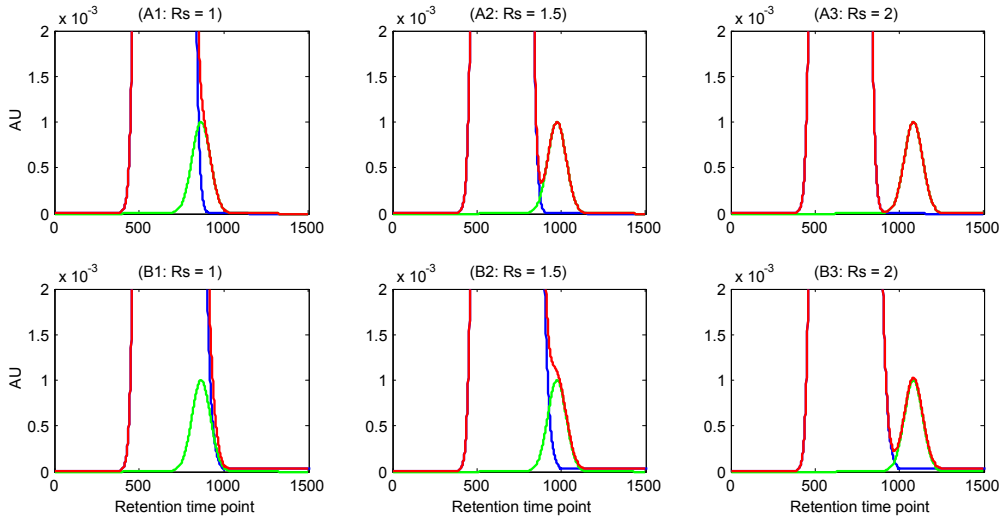

Figure 7. Different degrees of peak resolution. (A1 to A3): Symmetric target peak (blue) and 0.1% impurity peak (green) added together (red). (B1 to B3): Asymmetric (tailing) target peak and 0.1% impurity peak added together (red).

Common chromatographic practice often suggests that the minimum resolution between two peaks must be at least 1.5 to ensure sufficient separation. However, as illustrated in Figure 7 there is a remarkable difference in the actual peak separation depending on whether the target peak is symmetric or not. In Figure 7A2 the impurity peak is fairly separated

from the symmetric target peak at resolution 1.5, but in Figure 7B2 the impurity peak is partly hidden under the tailing edge of the asymmetric target peak. It is often difficult or impossible to detect such low resolution impurity peaks visually or to identify them by peak integration using existing commercial chromatographic 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 [40]. In PAPER II this challenge is addressed by monitoring the entire chromatographic profile both quantitatively and in a statistically reliable way. The automated multivariate statistical tool demonstrated in PAPER II is capable of detecting subtle changes in the chromatographic profile, specifically shoulders on the target peak as illustrated in Figure 7. These shoulders originate from small non-resolved impurity peaks, which would risk not to be detected by visual inspection and potentially be integrated as one peak using common generic peak detection and integration methods.

3.5 Peak purity

Detecting the occurrence of an unknown impurity co-eluting with the target compound is a particular problematic challenge. Therefore, purity analysis of a biopharmaceutical product often entails purity examination of the target peak. Peak-purity examination should prevent co-eluting impurities to escape detection in the conventional HPLC analysis [48]. HPLC with diode array detection (HPLC-DAD) is a commonly used method to conduct peak-purity examination. However, many impurities are structurally related to the drug substance, and their structure contains very similar chromophores, making purity assessment based solely on HPLC-DAD data difficult and unreliable. Coupling a mass spectrometer to a liquid chromatograph (LC-MS) brings more selective signals to the table. Since a mass spectrometer (MS) separates compounds by their respective mass-to-charge ratios (m/z), any difference in the m/z values between the impurities and the drug substance will allow an unambiguous detection regardless of similarities in their UV spectra. Therefore an impurity co-eluting with the target peak will be separated in MS as long as their m/z values are different and ionization of the impurity is not suppressed by the target compound. This is illustrated in Figure 8 where an insulin intermediate DesB30 is spiked with human insulin drug product at a 0.05% level [PAPER III]. Human insulin is co-eluting with

the structurally related target compound DesB30-insulin, but has a different molecular weight and thus different *m/z* values. The ion trace signals from human insulin have maximum intensity at *m/z* 1453. Plotting an extracted ion chromatogram (EIC) for this *m/z* value, the co-eluting profile of human insulin is provided (Figure 8).

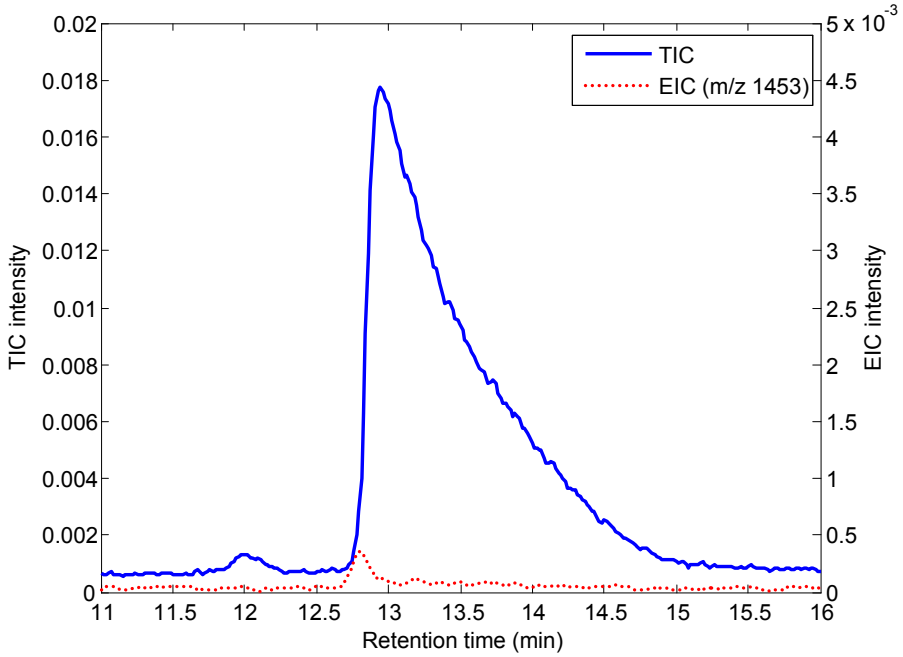

Figure 8. Plot of TIC and EIC (m/z 1453) of sample spiked with 0.05% HI [PAPER III].

It would be difficult or impossible to detect a co-eluting 0.05% impurity peak if measured with HPLC. However, with LC-MS this challenge is possible to meet and becomes practicable if assisted by automated multivariate statistical methods (described in Chapter 5).

4 Preprocessing

In the initial phase of the monitoring scheme applied in this study (subsection 2.3), historical chromatographic data are collected and preprocessed. In chromatography, the original data obtained from the instrument can be transformed into (possibly relative) concentrations of specific chemical analytes by integration of the peaks and the integrated data of selected peaks can then be monitored by multivariate statistical analysis [PAPER I]. However, this data reduction leads to a loss of information since the quality of the data relies on peak detection and on how the peaks are selected for the monitoring. Alternatively, the whole chromatographic data matrices collected from the instrument can be used, without first integrating a set of selected peaks [PAPER II and III]. The result is that not only peak magnitude is included but also its shape (peak overlap and peak shoulders). However, when monitoring whole chromatographic profiles or landscapes, instead of information on a limited set of peaks, some of the additional variation may obscure the relevant information. This extra unwanted variation is for example the variation originating from uninduced chemical variance, such as product sampling, sample work-up in the laboratory, and instrumental variation. For instance instrumental variation such as pressure, temperature and flow rate fluctuations may cause an analyte to elute at a different elution time in replicate runs. Additionally, matrix effects and stationary phase decomposition may also cause elution time shifting. Before multivariate statistical monitoring can be performed, the data should be corrected for this unwanted variation, since it largely influences the outcome of the monitoring and disturbs monitoring of the *chemical* variation. Using mathematical preprocessing methods, this unwanted variation can be removed or handled. Several methods can be applied to prepare the chromatographic signal for subsequent multivariate statistical monitoring. So far only few preprocessing methods are implemented in commercial chromatographic software and they often tend to be too simple and generic. Here the focus will be on selected preprocessing methods that have found their usefulness for the applications described in PAPER I, II, and III. The preprocessing methods are described in the order that they preferably should be applied to chromatographic data.

4.1 Baseline correction

Baseline correction in chromatography is commonly employed. Baseline variation has been an issue in chromatography for decades, and one of the first descriptions on how to remove baseline drifts was presented already in 1965 [49]. Nowadays most methods are based on subtracting a fitted polynomial following the baseline curvature, and several such methods are available in the literature [50,51]. Among the different approaches of baseline correction, this thesis favors an approach proposed by van den Berg [52]. This baseline correction method operates in local regions of the chromatogram and uses B-splines constructed from polynomial pieces joined at certain positions (knots). The method operates by gradually eliminating points in the signal furthest (northern distance) away from the fitted polynomial until the number of selected supporting 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. In PAPER II the baseline correction method by van den Berg [52] was applied even though only minor baseline drifts were observed. Nevertheless, the developed monitoring approach should be capable of handling more severe baseline drifts if such appear. The baseline correction is illustrated in Figure 9 (modified data from PAPER II), where a chromatogram with minor (A + B) and major (C + D) baseline drift is corrected, using the same settings.

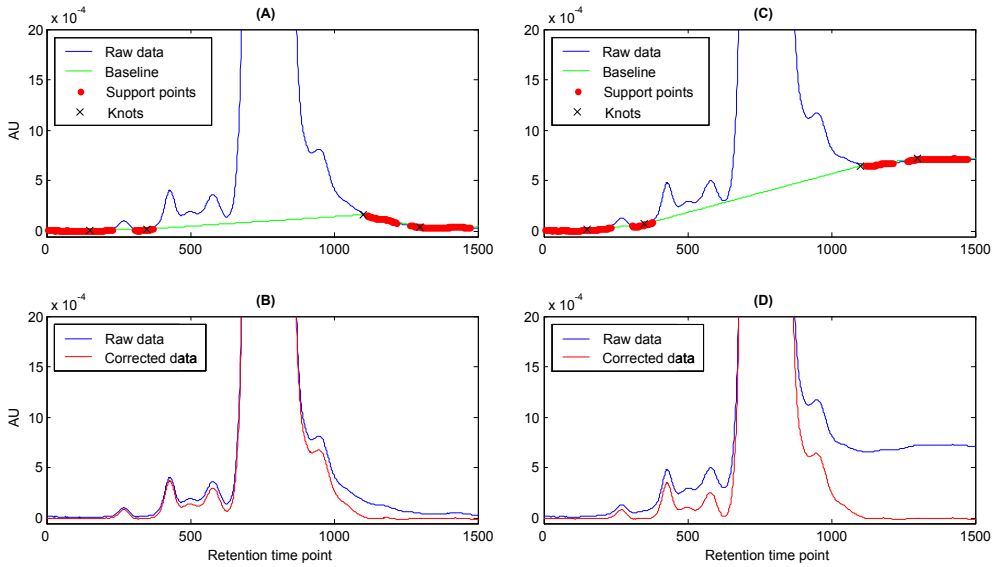

Figure 9. Illustration of baseline correction method by van den Berg (2008) [52]. (A) Raw data with minor baseline drift, knot positions, and fitted baseline between knot positions. (B) Data with minor baseline drift before and after baseline correction. (C) Raw data with major baseline drift, knot positions, and fitted baseline between knot positions. (D) Data with major baseline drift before and after baseline correction.

By inspection of Figure 9 it can be confirmed that the baseline correction is capable of handling various degrees of baseline drifts using the same settings. Thus, upon selecting the settings from initial data investigation, baseline correction can be an objective and automatic preprocessing step.

For LC-MS data a variety of techniques for baseline correction are applicable and these are reviewed by Listgarten et al. [53] among others. In PAPER III an efficient and rather simple method for baseline correction was applied. The method works by fitting a global polynomial (of a user-defined order) to each extracted ion chromatogram of the LC-MS landscape and, through an iterative routine, down-weighting points belonging to the signal. A baseline is then constructed and subtracted from the original extracted ion chromatogram [PAPER III]. The baseline correction method is similar to a previously described method by Gan et al. [51].

4.2 Normalization

Normalization of chromatographic data is another possible step in the preprocessing procedure. Normalization is a sample-wise standardization of

data, usually applied to remove a source of unwanted variation. In chromatography it is common to apply normalization to minimize the effect of variation of sample size that actually hits the column, possible sample carry-over, and drifts in e.g. detector efficiencies. Normalization procedures enable a more accurate matching and quantification between multiple samples. Different procedures for normalization can be applied, such as setting maximum peak height to the same value for all samples, or dividing each signal value for one sample by the sum, mean, or median of all signal values for that sample. In Figure 10 the effect of different normalization procedures are illustrated on LC-MS data (modified from PAPER III).

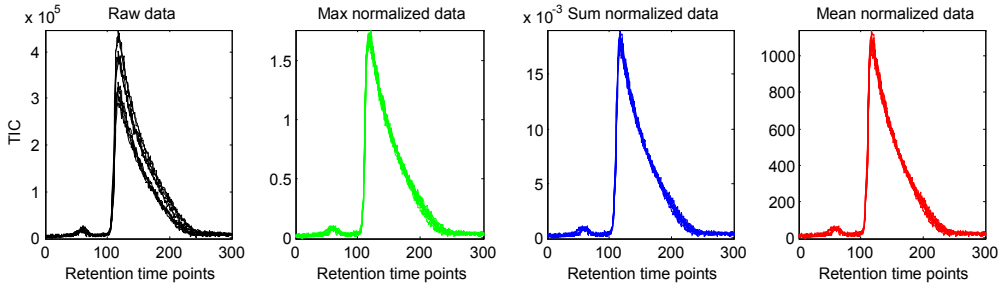

Figure 10. Illustration of raw LC-MS TIC data and the effect of different normalization procedures (modified from PAPER III).

The different normalization procedures illustrated in Figure 10 all correct for the bias between samples in the raw data. The only real difference between the procedures is the scale to which the data are normalized. For the application described in PAPER III it was assumed that the target peak purity might vary but the overall signal intensity should ideally be the same for each sample. Therefore the sum of all intensities was used as normalization value for each sample.

Although normalizing the data generally improves comparison of samples across instrument runs, the applied approaches are independent of or "blind" to the actual compound level in the sample. Ideally, the use of spiking controls would be an appropriate option for addressing the instrumental variability. However, this approach is rather laborious.

4.3 Alignment

As with every laboratory experiment, chromatographic separation is stable and reproducible only to a certain extent. The retention time often shows

large shifts, and distortions of elution profiles can be observed when different runs are compared. For LC-MS data even the MS (m/z) dimension might show (typically smaller) deviations. Alignment of shifted peaks can be performed in various ways. During the past decades, several kinds of useful alignment approaches have been developed for chromatographic profiles [50,54-62]. 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 so-called correlation-shifting (coshift) algorithm [63] or the faster interval-correlation-shifting (*i*coshift) algorithm [64] described in subsection 4.3.1. Yet, if the column is changed between runs or if samples are measured over a long period of time, this may cause the peaks to shift independently from one another in the same chromatogram, and more complex shift correction is needed to correct for this non-systematic shift. One of the most popular and efficient methods, which can handle this non-systematic shifts in chromatographic data, is the piecewise alignment algorithm correlation optimized warping (COW) [57,61] described in subsection 4.3.2. The algorithms mentioned here all use a target or reference chromatogram that each chromatogram is aligned towards. The choice of reference chromatogram is an important aspect of the alignment methods considered here, and will be described further in subsection 4.3.3.

4.3.1 Systematic shift correction by interval-correlation-shifting

The *i*coshift algorithm is originally developed for alignment of nuclear magnetic resonance (NMR) spectra [64], but has proven to be well suited for alignment of chromatographic data as well [PAPER II and III]. The algorithm independently aligns each chromatogram to a reference by maximizing the cross-correlation between user-defined intervals and employs a fast Fourier transform engine that aligns all chromatograms simultaneously. The *i*coshift algorithm is demonstrated to be faster than similar methods found in the literature making full-resolution alignment of large datasets feasible [64]. Several options are available depending on how the alignment problem is to be solved. For instance, it is possible to define intervals to be aligned separately (e.g., allowing a full chromatographic alignment with regularly spaced intervals, or adjacent intervals of user-defined length, or customized interval boundaries). Furthermore it is possible to set a boundary for the maximum local correction allowed for each interval. Finally, the fill-in value is defined for the reconstruction part

(a missing value or the first/last point in the interval). In Figure 11 the result of *i*coshift alignment of a shifted chromatographic profile (one interval) towards a reference is illustrated (modified from PAPER II).

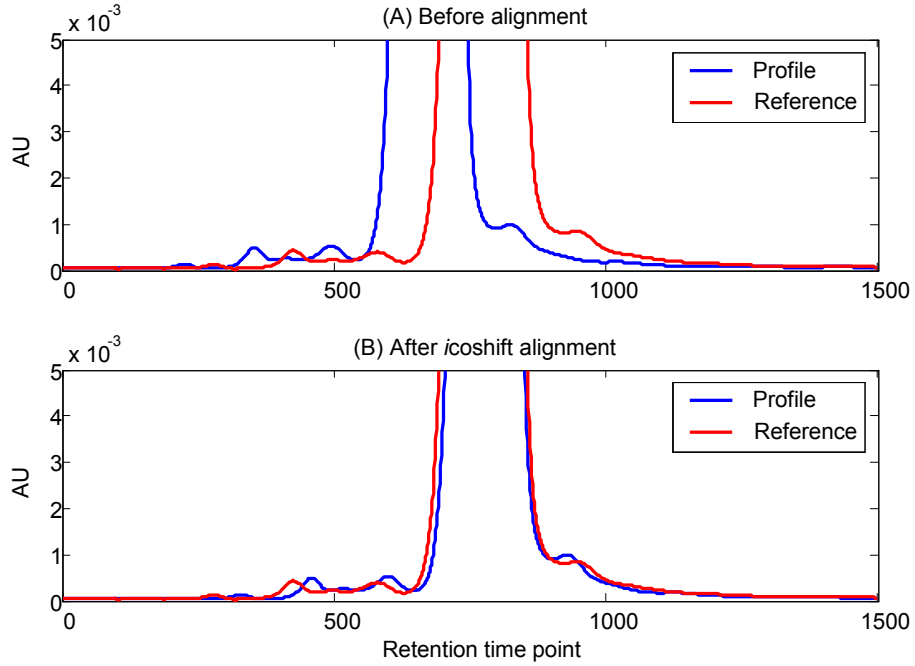

Figure 11. Alignment of a profile chromatogram (blue) towards a reference chromatogram (red). Chromatograms before (A) and after (B) alignment using the *i*coshift algorithm (modified from PAPER II).

The *i*coshift alignment illustrated in Figure 11 clearly handles the major systematic shift. However, some non-systematic shifts still remain uncorrected, especially for the minor peaks. To solve this more complex shift correction is needed. Even though the *i*coshift alignment could not correct for the entire retention time shift, it still serves a purpose. Most often both a preliminary systematic shift correction is needed before a non-systematic shift correction can be handled successfully.

4.3.2 Non-systematic shift correction by correlation optimized warping

The Correlation Optimized Warping technique (COW) was originally introduced by Nielsen et al. [57] as a method to correct for shifts in vectorized data signals. COW is a piecewise or segmented data alignment technique that uses dynamic programming to align a sample chromatogram

towards a reference chromatogram by stretching or compression of sample segments using linear interpolation [61,62]. Two input parameters are required and the first parameter is a number of sections into which the chromatograms is divided (by knots). The second parameter, the so-called warping parameter, defines the degree of alignment (slack). For the larger values of the warping parameter the larger time shifts can be corrected [42]. The performance of COW alignment is illustrated on the chromatographic profile after *i*coshift alignment as described in subsection 4.3.1. Here the profile was divided into four sections as indicated by the knots.

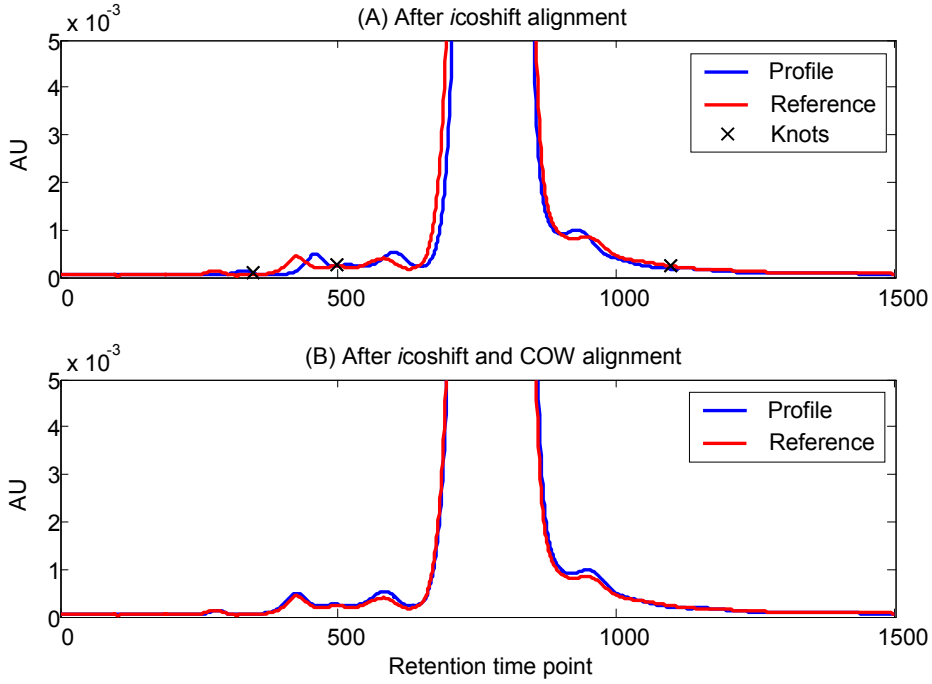

Figure 12. Alignment of a profile chromatogram (blue) towards a reference chromatogram (red) divided into four sections indicated by the knots. Chromatograms after *i*coshift alignment (A) and after *i*coshift and COW alignment (B) (modified from PAPER II).

As illustrated in Figure 12 the COW algorithm offers much better alignment, but the selection of sections and the warping parameter is crucial. The computational time of COW is exponentially influenced by the warping parameter. If the alignment is unsatisfactory, more sections or a larger warping parameter value can be considered. However, it is often possible to

achieve good alignment at a low warping parameter, thus ensuring reasonable computation time [42].

4.3.3 Selection of reference chromatogram

Several methods can be used for finding a proper reference chromatogram for alignment. Among these are, the average chromatogram, the first loading of a PCA model, the most inter-similar or representative chromatogram containing the highest number of common peaks [61,65,66], or the sample run in the middle of a sequence [59,67]. Furthermore, using the chromatogram with the highest correlation coefficient with respect to the remaining chromatograms as reference has also been suggested [62,68]. This approach was favored in both PAPER II and PAPER II. 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 [62,68].

4.4 Data reduction

To make chromatographic samples even more comparable, data reduction (binning or bucketing) can be applied. Binning can be performed in various ways, e.g. by summing or averaging all intensities within a user-specified bin level. This may reduce small uncorrected chromatographic artifacts, such as shifts. Furthermore binning simplifies subsequent multivariate statistical analysis, as the huge amount of data points per sample is reduced.

In PAPER III the data reduction puts all the intensities on a (time, m/z) grid and sum the intensities within each bin. The effect of LC-MS data reduction using different bin sizes is illustrated in Figure 13.

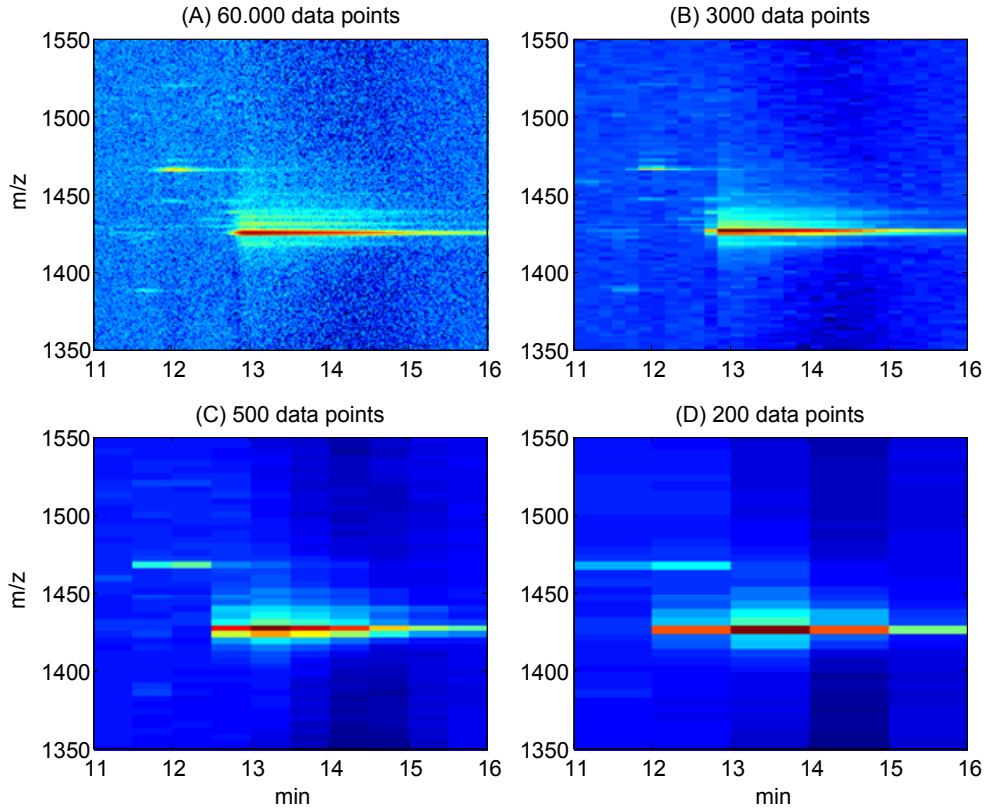

Figure 13. Illustration of LC-MS data reduction using different bin sizes. (A) Before data reduction (60.000 data points); (B) bin size: 10 seconds and 2 *m*/*z* (3000 data points); (C) bin size: 30 seconds and 4 *m*/*z* (500 data points); (D) bin size: 60 seconds and 5 *m*/*z* (200 data points (modified from PAPER III).

The bin size should be selected based on experience and the sample being tested. However, in PAPER III the optimal bin size was selected based on testing. The impurity detection level was tested using different selections of bin size and consequently bin number. The lowest detection level was obtained with a bin size from 30 to 60 seconds and 1 to 2 *m*/*z* value, resulting in 500 to 2000 bins [PAPER III].

4.5 Scaling

Scaling is a variable-wise standardization and the choice of scaling method is crucial for performance of the subsequent multivariate statistical monitoring. Scaling methods 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 (i.e. to bring all variables into the same range), and

to correct for a non-constant signal variance. For instance a fold difference in concentration for the target compound and a related impurity may not be proportional to the chemical relevance of these compounds [69].

Mean centering may solve this problem by subtracting the average variable pattern from each sample. This removes a common offset, and brings each variable to vary around zero. However, mean centering may not always be sufficient, hence autoscaling, also referred to as unit variance scaling, can solve the problem by dividing all mean centered numbers of a variable by the standard deviation of that variable [70]. After autoscaling, all variables have mean values zero and a standard deviation of one. Therefore the data is analyzed on the basis of correlations instead of covariances, as is the case with mean centering [71]. The effect of mean centering and autoscaling is illustrated (Figure 14) on a peak table dataset (*15 samples × 20 peak areas*) obtained from integration of one major target peak and nineteen related minor peaks (modified from PAPER I).

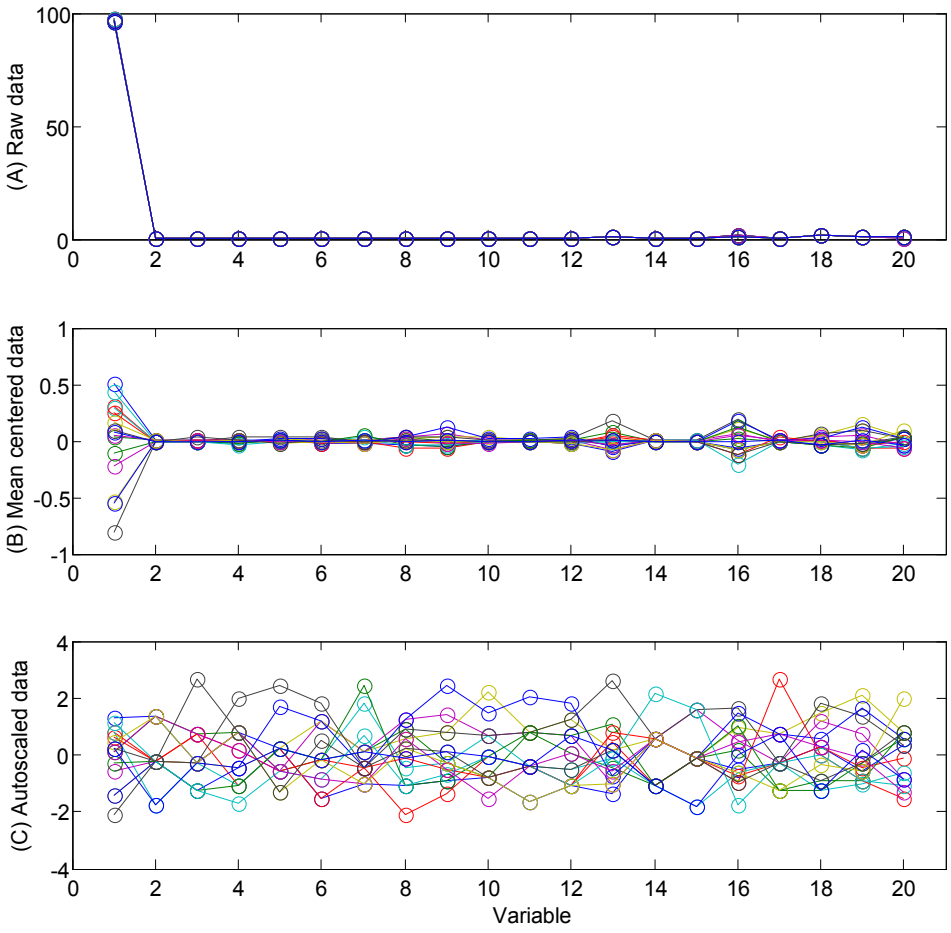

Figure 14. Illustration of a peak table dataset (*15 samples ×20 peak areas*) before (A) and after mean centering (B) and autoscaling (C) (modified from PAPER I).

After mean centering (Figure 14B), all variables will have mean values zero. Mean centering is normally recommended for data where the variables have same units. After autoscaling (Figure 14C), all variables have equal length and mean values zero. Autoscaling is recommended for data where the variables have different units or if the variation in range of different variables is large. For the example illustrated in Figure 14 (modified from PAPER I), variable 1 (target peak) originates from the high concentration target compound, with large absolute fluctuations between samples. This is not desired since the other related compounds giving rise to smaller peaks and peak variation are equally interesting for the application in PAPER I. When processing full chromatographic profiles, however, the use of autoscaling magnifies the baseline variation since variables (i.e. retention

time points) representing only noise will also be transformed to the same scale as all the other variables (Figure 15B). Baseline variation will thus become equally important as variation in chromatographic peaks. This also holds true for components with very low concentrations and variation. One effective way to reduce the relative importance of large values without blowing up noise is square root mean scaling (Figure 15C). This scaling method uses the square root of the mean (of individual variables) as scaling factor. In PAPER II square root mean scaling turned out to be the most appropriate scaling method, as it first of all increased the sensitivity on detecting small unknown peaks partly hidden under the target peak. Secondly, the characteristic appearance of the chromatogram was kept intact, which was helpful when interpreting a faulty chromatogram detected by the multivariate statistical model.

Scaling may also be crucial in order to bring the distribution of data points close to a normal distribution. This is especially important when multiple testing (like Student's t-test) is used for difference analysis as in PAPER III. In many cases, a logarithmic transformation is used for stabilization of the variance. Logarithmic transformation can also be a solution to the problem of difference in range of variables; however, it may result in the same noise drawback as autoscaling when processing full chromatographic profiles (Figure 15D). Furthermore, log transformation of zeros and negative values is a problem. The different scaling methods mentioned here were applied to fifty chromatographic profiles and the effect is illustrated for one profile in Figure 15 (modified from PAPER II).

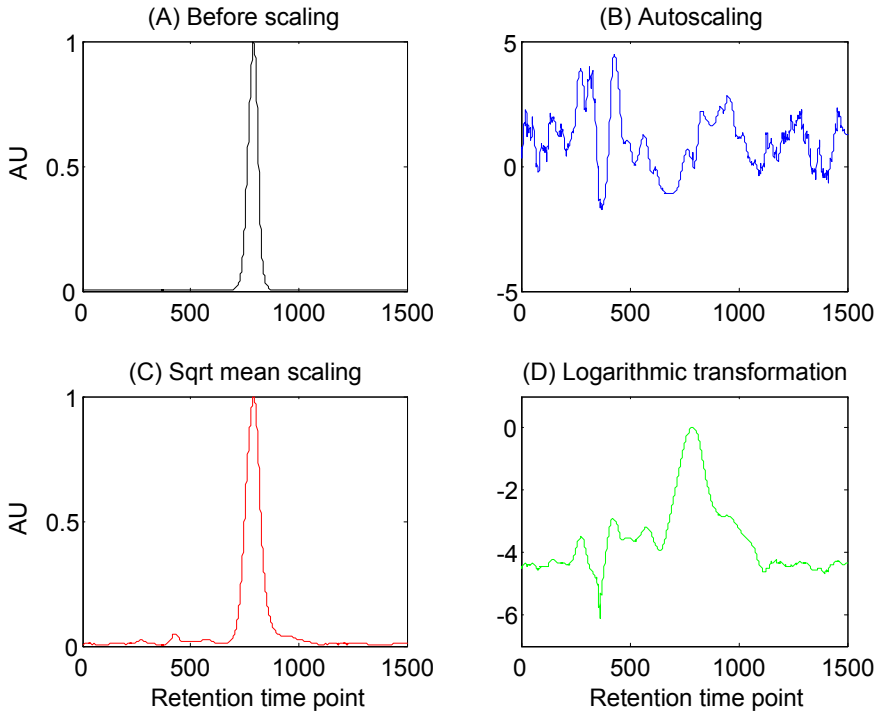

Figure 15. Effect of different scaling methods applied to fifty chromatographic profiles. Here one profile is plotted before scaling (A), after autoscaling (B), after square root mean scaling (C), and after logarithmic (log10) transformation (D) (modified from PAPER II).

There are several other centering, scaling and transformation methods which not are mentioned here, some of them are well described in a paper by van den Berg et al. [69].

5 Multivariate statistical monitoring

This chapter covers all the elements in the training phase and application phase of the monitoring scheme (subsection 2.3). These include modeling of preprocessed normal operation condition (NOC) chromatographic data, construction of multivariate control charts, monitoring of new chromatographic data, and diagnosis of deviating chromatographic data using contribution plots.

To be able to explain the multivariate statistical monitoring to a broad audience, the confusion between process monitoring and process control needs to be clarified. From a chemical engineering point of view, process control is about automated surveillance with well-defined control actions of a process. However, in multivariate statistical process control (MSPC) the 'normality' of the process is statistically determined and monitored. The underlying concept of MSPC is based on a comparison of what is happening today with what happened previously. Hence, MSPC is actually a technique for statistical monitoring of processes in spite of the fact that the designation suggests that actual control actions are performed. Therefore, MSPC is referred to as multivariate statistical monitoring in the ongoing to avoid misunderstandings.

This work advocates the use of latent variables-based MSPC [72], specifically MSPC based on principal component analysis (PCA) [73]. The PCA-based MSPC approach developed here, considers all the noisy and highly correlated chromatographic variables, but project this information down onto low dimensional subspaces which contain the relevant information. The chromatographic data is then monitored in this latent subspace by using a few multivariate control charts built from multivariate statistics. PCA-based MSPC is suitable for monitoring two-way data, such as tables of peak areas or chromatographic profiles where each sample is a vector of values collected in a data matrix. However, MSPC based on PCA cannot handle so-called three-way chromatographic data structures, such as LC-MS data, where each sample is a matrix of values collected in a data cube or tensor. If PCA-based MSPC should be applied to LC-MS data, then one dimension must be reduced either by summing or unfolding (see subsection 3.2). Alternatively MSPC based on multiway methods, such as PARAFAC [43]

(an extension of PCA to multiway data) or PARAFAC2 [74] (handles retention time shifts), can handle the additional mass spectral dimension of LC-MS data without reducing the three-way structure. These multiway methods give new possibilities with regard to the information that can be extracted, but are not as widespread and user-friendly as two-way methods, due to their more sophisticated model nature. In situations where both process variables and product quality data are available, multivariate predictive models based on projection to latent structures like partial least squares (PLS) [75] can be applied. These multivariate predictive models can be used to develop a predictive relationship between the process variables and the product quality (if present). In this way PLS-based MSPC can monitor the measured process variables and from these estimate the product quality. However, none of these multiway or predictive methods have found their use in this PhD project, and are therefore out of the scope of this thesis.

This chapter will include different aspects of MSPC based on PCA in chromatography. The usefulness of bringing PCA-based MSPC and chromatographic data together will be discussed and examples will be given from the papers prepared during this thesis. For chromatographic data, PCA-based MSPC can either be applied to integrated peak areas in peak tables (discrete data), or to fingerprints or whole chromatographic profiles (continuous data). The first paper [PAPER I] in this thesis deals with multivariate statistical monitoring of peak tables, whereas PAPER II and PAPER III cover the monitoring of chromatographic fingerprints. In the previous chapter it was described how preprocessing of the chromatographic data can be applied to generate 'cleaner' data so relevant variation is more predominant in the data. Preprocessing is the first step and a prerequisite for monitoring relevant information. For the following discussion, it is assumed that the chromatographic data is properly pre-processed.

Since the theory of MSPC originates from statistical process control (SPC), it is relevant to give a brief introduction to the concepts of SPC and the link to MSPC.

5.1 Statistical process control

The basic idea of statistical process control (SPC) is to monitor the performance of a process over time in a so-called control chart. The greatest developments of statistical process control have taken place during the 20th

century. In the 1920's statistical theory began to be applied effectively to quality control as a result of the development of sampling theory. W.A. Shewhart [76] was the first to develop and describe the fundamentals of SPC in the early 1930's, and the control chart found widespread use during World War II and has been employed, with various modifications ever since. Shewhart's work pointed out the importance of reducing variation in a manufacturing process for improvement of the end-product quality. The process variation can be well monitored with the use of control charts, which eventually leads to adjustments of the process. Shewhart distinguishes between variation that is normally expected of the process due *chance* or *common-causes* (the usual, historical, quantifiable variation in a system), and variation that changes over time due to *assignable* or *special-causes* (unusual, not previously observed, non-quantifiable variation) [17].

The Shewhart $\overline{X}$ -chart (from now on referred to as control chart) is a simple plot of the quality characteristic vs. time or sample. The control chart makes assumptions about the data, namely that it is independent, and it is normally distributed. Usually the control chart consists of a centerline (mean value), two warning limits (mean ± 2σ), and two control limits (mean ± 3σ) which indicate the range of variation of the quality characteristic (Figure 16).

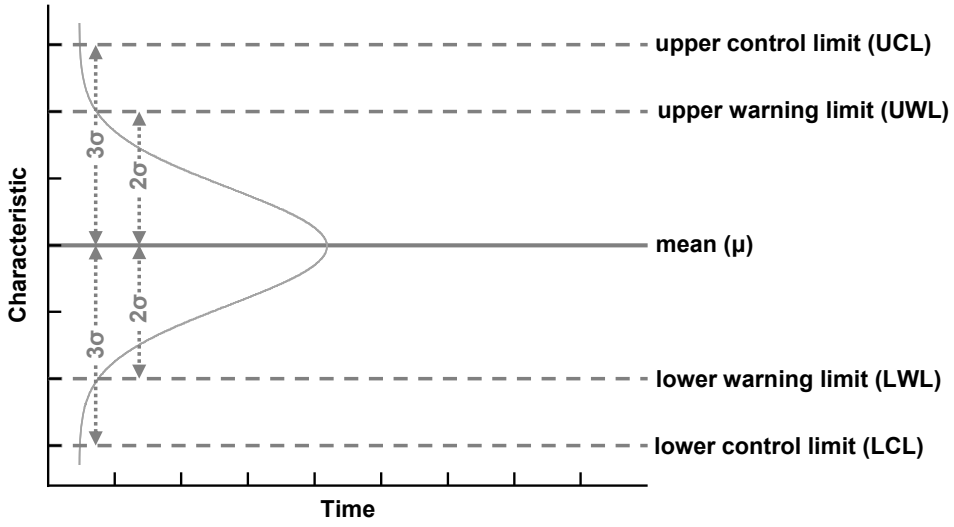

Figure 16. Control chart with warning and control limits.

Different so-called run-rules help interpret the control chart in order to distinguish between out-of-control and in-control situations. However, the most important rule and a basic criterion is that one or more observations

outside of the control limits is considered rare, and indicates that the variation is due to an assignable cause and the process is out-of-statistical control.

Alternative control charts have been developed to detect small shifts of the mean. These are e.g. the CUSUM chart [77,78] and the EWMA chart [79]. These alternative control charts will not be explained here.

Most SPC approaches are based upon the control charting of a small number of variables, and examining them one at a time (univariate). This is inappropriate for many process applications where several variables of importance are available. The practitioner cannot reliably study more than two or three charts to maintain overview of the process. The risk of making mistakes is larger when many control charts are to be checked [17]. Furthermore, the univariate control charts do not account for the correlation structure in the data. This is exemplified in Figure 17.

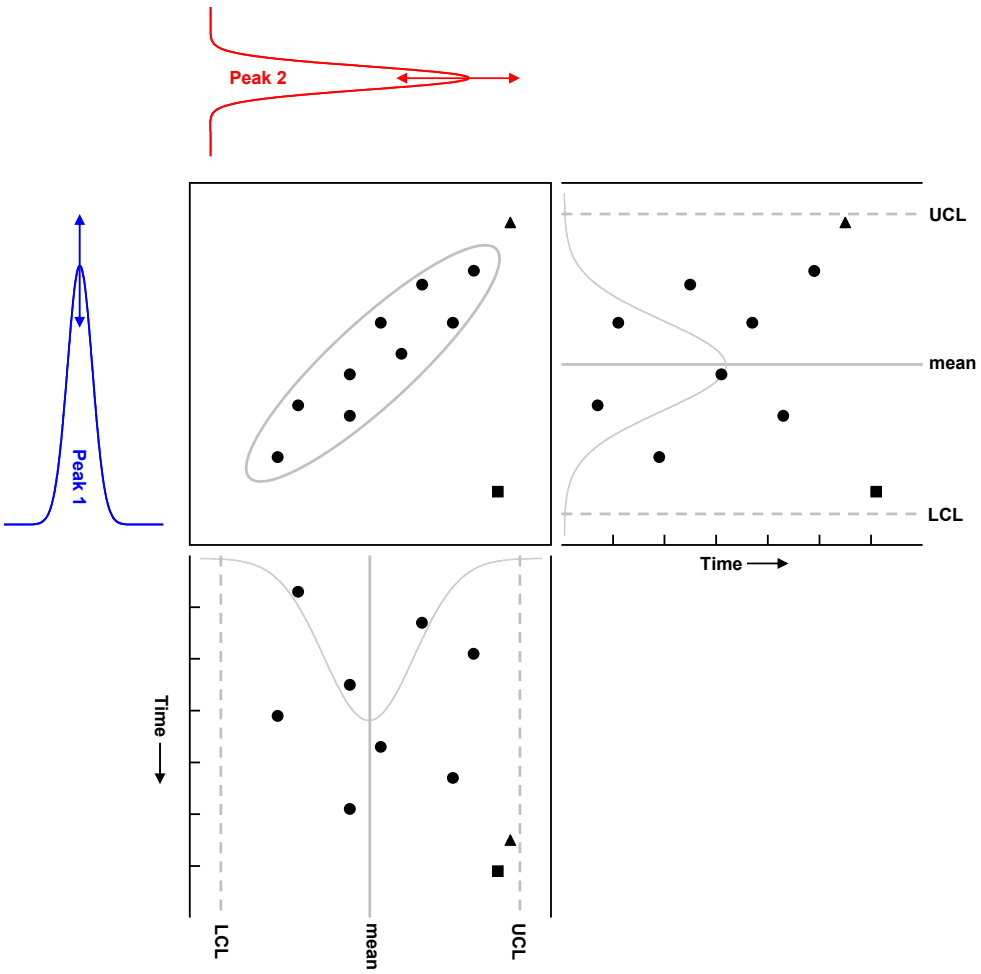

Figure 17. Outline of different situations in univariate control chart monitoring (modified from Nijhuis et al. [80]).

In Figure 17 a two-dimensional data set composed of the areas of two chromatographic peaks is presented in both a univariate- and a multivariate way. The ellipse in the scatterplot represents the correlation structure in the data. In order to compare the univariate statistical approach with the multivariate approach, the univariate control charts of peak 1 and peak 2 are given. All nine ● observations are describing *common-cause* variation both in a univariate- and a multivariate sense. The ▲ observation does not deviate from the correlation structure but is clearly an extreme both in a univariate and a multivariate sense. The ■ observation seems to be within common-cause variation in a univariate sense, but clearly deviates in a multivariate sense. This is caused by the fact that the ■ observation departs from the

correlation structure in the data. The univariate charts are clearly missing a faulty situation as a consequence of the correlation structure in the data which is not taken into account. The principle of multivariate control charts is of course of more interest when one has to deal with a higher dimensional data set.

5.2 *Multivariate statistical process control*

Contrary to univariate SPC which typically deals with single observations, multivariate SPC (MSPC) techniques can handle many and correlated variables. This is often relevant in industrial processes, where relationships between the variables have to be taken into account. Univariate control charts applied to multivariate systems are often inadequate at detecting and handling a fault or an abnormality in the operation. This is because the process variables often are correlated, and a special cause can affect more than one variable at the same time. MSPC takes this correlation into account in monitoring the mean vector or variance-covariance matrix. In MSPC, historical data are used to calculate empirical statistical models that describe the acceptable trends of the whole system, using latent variables instead of every measured variable. When a problem appears, it changes the covariance structure of the model and this can be detected using multivariate statistics.

Hotelling [81] was one of the first who introduced a multivariate approach for SPC of a process in the 1940's. He applied his procedures to bombsight data during World War II. In the late 1950's Jackson [82] applied principal component analysis (PCA) to reduce the dimensionality of several related variables and introduced the control chart for T^2 of principal components. In the late 1970's Jackson and Mudholkar [83] investigated PCA as a tool of MSPC and introduced a residual analysis. The control chart was introduced for the sum of squared residuals Q as well as T^2 of principal components retained in a PCA model. In the early 1990's the main concepts behind the development and use of latent variable-based multivariate SPC for monitoring continuous processes were provided by Kresta et al. [84], Wise et al. [85], Kourti and MacGregor [86]. Illustrations of the methods along with the algorithms and details on estimating control limits are well described by Kourti [87,88], Montgomery [17], Bersimis et al. [89], Ferrer [21], and Westerhuis et al. [90].

Nijhuis et al. [91] were some of the first to apply MSPC in chromatography in the late 1990's. Since then, there are several examples where a PCA based MSPC scheme has found its use in the chromatographic discipline [92,93]. These were all based on monitoring the analytical signals, i.e. peak areas or peak tables. Additionally, MSPC has also been applied for the surveillance of chromatographic instrument systems where instrument-related parameters were modeled [94,95]. Here focus is thus not only put on post run checks of peak areas but rather on monitoring of the analytical process itself.

In the following, the concepts of MSPC based on PCA will be explained using in terms of analytical chromatographic signal data.

5.3 PCA

Principal component analysis (PCA) is a common technique used for dimensionality reduction and is implemented in all multivariate data analysis software packages and also in some instrument software. PCA was originally developed by Pearson in 1901 [96], though it is more often attributed to Hotellings work from 1933 [73], where he described and developed PCA to its present stage. Since then PCA has been used for several applications in different scientific disciplines, amongst others in the area of chemometrics (defined as the application of mathematical and statistical methods to chemical measurements [97]). Comprehensive information on the principles and applications of PCA can be found in several good reviews [97-99] and text books [71,100].

PCA is a bilinear model that finds combinations of variables that describe common patterns in a given data set $\mathbf{X}$ ($M \times N$) with M rows of samples and N columns of variables. Mathematically, PCA is based on decomposition of the covariance or correlation matrix of the variables along the directions that explain the maximum variation of the data. The matrix $\mathbf{X}$ can be decomposed by either the NIPALS or the singular value decomposition (SVD) algorithm [98]. For a given data matrix $\mathbf{X}$ the covariance matrix of $\mathbf{X}$ is defined as:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{M-1} \tag{4}$$

This assumes that the matrix $\mathbf{X}$ has been mean centered (i.e. adjusted to have zero mean by subtracting the mean of each column). If the columns of $\mathbf{X}$ have been autoscaled (i.e. adjusted to zero mean and unit variance by

dividing each column by its standard deviation), Equation 4 gives the correlation matrix of **X** [101]. PCA establishes new directions in the original data cloud; so-called latent variables or loadings (**P**), which are constructed as linear combinations of the original variables. The first new direction is found so that the maximum variance in the original data is explained. For the first direction, each sample (from its original position) can be projected onto this, providing a score value (**T**). These score values then describe the amount of the latent variable/loading found in each sample, whereas the loadings contain information on how variables relate to each other. A set of a score and loading vector constitutes what is denoted a principal component ($\mathbf{tp}^T$) or PC. The direction of a principal component in relation to the original variables is given by the cosine of the angles α_1, α_2, and α_3 (loading coefficients) as illustrated in (Figure 18). The second new direction in the data is found orthogonal (the mathematical constraint used for PCA) with respect to the first direction and the second score value for each sample is found in a similar fashion as described above. This is continued as long as systematic (descriptive) variation is described by the successive principal components. The variance explained in each principal component decreases for successive extracted components.

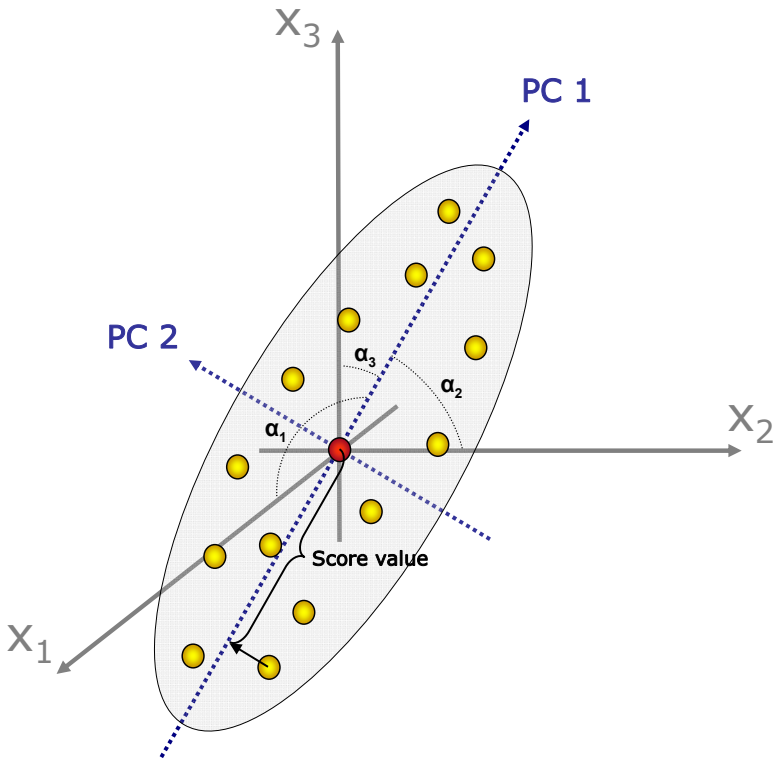

Figure 18. Illustration of the first and second PC, representing maximum variation in the mean centered data (yellow circles). Both PC1 and PC2 are passing through the average point (red circle). Each sample may be projected onto the PC to get a score value. The direction of PC1 in relation to the original variables is given by the cosine of the angles α_1, α_2, and α_3 (loading coefficients for PC1).

Figure 18 shows the new axes PC1 and PC2 created by PCA. There is greater variance on the PC1 axis than on the PC2 axis. This is not surprising as PC1 is constructed to lie on the direction of the greatest variance in the data. By ignoring higher-order components, a new version of the data with fewer variables than the original data is generated. The variance left in the data (unexplained variance) is usually related to unsystematic variation or noise and is termed the residuals (**E**). Mathematically, PCA decomposes the data matrix **X** as the sum of the outer product of the score vectors $\mathbf{t}_i$ and the loading vectors $\mathbf{p}_i$ plus a residual matrix **E**:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E} = \mathbf{T}\mathbf{P}^\mathbf{T} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \qquad (5)$$

where **T** ($M \times R$) is the score matrix and **P** ($N \times R$) is the loading matrix, with R components. Here R must be less than or equal to the smaller dimension of **X**, i.e. $R <= \min(M,N)$. $\hat{\mathbf{X}}$ is the matrix of predicted or reconstructed values.

Applications of PCA rely on its ability to reduce the dimensionality of the data matrix while capturing the underlying relationship between the variables. To illustrate this from chromatography a simple two-peak system is depicted in Figure 19:

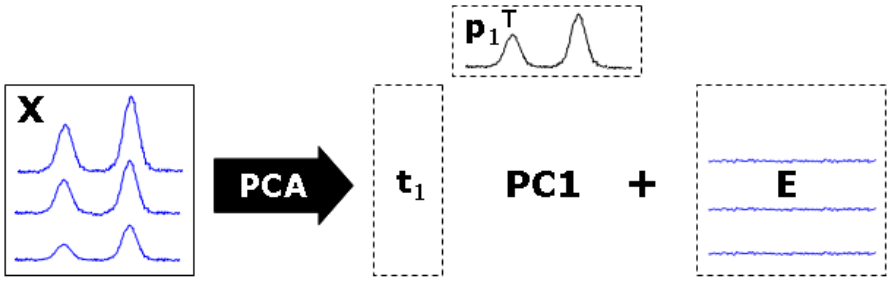

Figure 19. Illustratrion of Principal Component Analysis (PCA) of a simple chromatographic two-peak system for three samples. X is the original data matrix, p_1 is the first loading vector (common profile) and t_1 the score vector holding the amount of the first loading. No noise is present in the data and thus one principal component (PC1) will explain all variance; i.e. the residual matrix E is zero (modified from PAPER II).

The two-peak chromatographic profiles depicted in Figure 19 are rather simple, as the only difference between the three samples is the peak heights, i.e. the ratio between the two peaks, and thus the chromatographic profile, is the same for all three samples. The PCA model captures the maximum variation, which follows the chromatographic profile and therefore, the first loading resembles the original data. The score value is then simply a measure of the magnitude of the chromatographic profile and can be used as a direct measure of relative peak area or concentration. Because PCA is a bilinear model, twice as high concentration (peak area) gives twice as high a score value (assuming no noise and baseline is present and similar peak shape regardless of the concentration).

If the chromatographic profile also varies between the samples, a more comprehensive PCA model is needed. This is illustrated in Figure 20.

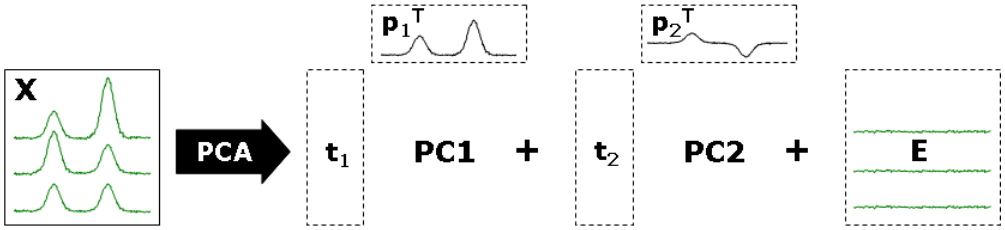

Figure 20. Illustratrion of PCA of a chromatographic two-peak system for three samples. X is the original data matrix, p_1 and p_2 is the first loading vectors (common profiles). t_1 and t_2 are the score vectors holding the amount of the first two loadings respectively. No noise is present in the data and thus two principal components (PC1 and PC2) will explain all variance; i.e. the residual matrix E is zero (modified from PAPER II).

In Figure 20 both the peak heights and the peak ratio varies between the three samples. The PCA model includes this variation by using two principal components to describe both the magnitude and the different chromatographic profiles. Now the two loadings together describe the common profile.

The correct number of significant principal components can be determined in several ways. An obvious method is to select those components that together account for a large percentage of the total variance captured. In addition, the inspection of loadings can verify whether the components seem to reflect any clear systematic variation or just noise. This is exemplified in Figure 21, where the first two loadings, from a PCA model on unfolded LC-MS data, is folded back and plotted in 3D to help interpretation [PAPER III]. The inspection of loadings confirmed that the first two components reflect real systematic chemical variation.

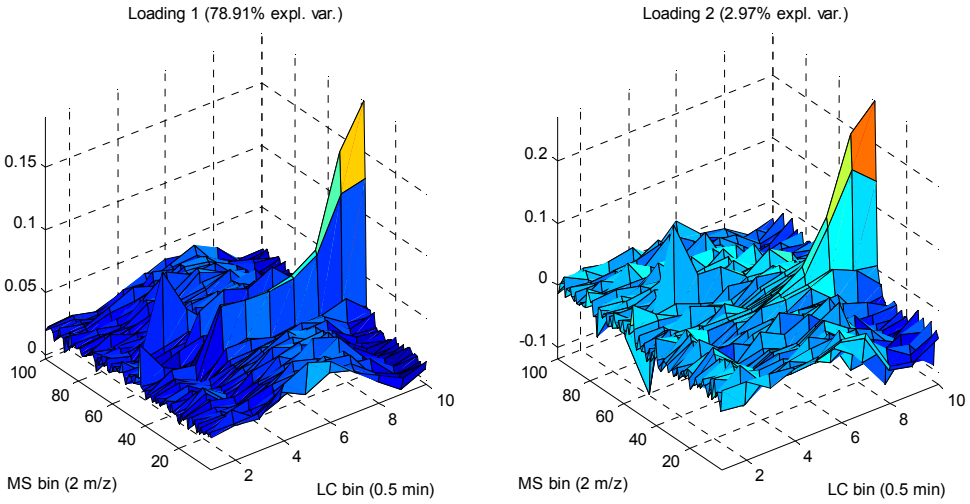

Figure 21. 3D plot of the first two PCA loadings [PAPER III].

Another possibility to determine the correct number of principal components is the evaluation of the prediction error during validation, which is described in subsection 5.4.

Once the samples in the data matrix (**X**) have been modeled with PCA, new samples ($\boldsymbol{x}_{new}$) can be fitted to the model. This is illustrated in Figure 22.

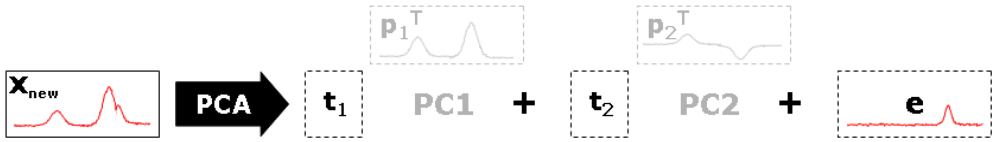

Figure 22. Prediction of a new sample (x_{new}) using the model loadings to generate new score values (t_1 and t_2) and residuals (e) (modified from PAPER II).

In Figure 22 a new sample is predicted using the already defined model loadings, and consequently new score values and residuals are generated. Most of the chromatographic profile for the new sample is described by the model loadings. However, the small third peak in the new sample is not described by the components retained in the model. Accordingly the third peak shows up as an abnormal residual variability. This information can be utilized when monitoring the chromatographic profile, as published in PAPER II.

5.4 Validation

An important element in multivariate statistical analysis is the validation of the calculated models. This is to avoid false correlations, determine the optimal number of components to use in the model and to ensure that the estimated model reflects reality. The integrity and applicability of the derived model are totally dependent on the set of data used to build the model. Hence, model validation is a critical aspect to ensure that the model is representative of the variations to be encountered in future samples.

Cross-validation [102] is an internal re-sampling method and the most often used method for error estimation in PCA. In cross-validation new data sets are created by systematically removing samples from the data set, either in small segments or individually as in leave-one-out cross-validation. The residuals for the samples that were left out, using the model built with the remaining samples, serve as a measure for the overall prediction error. Cross-validation is typically used in preliminary studies of data and if a data set is limited to very few objects (less than 50). More details on the most commonly used generic PCA cross-validation methods can be found in a review by Bro el al. [103].

Test set validation is a method for validation of a model by another data set, which can be either dependent or independent. A truly independent data set represents a separate selection from the entire population e.g. samples collected from another period of time where all possible sampling errors and sample variations are present. This kind of validation is the ultimate validation of any model and is also referred to as external validation. If an independent test-set is not available, the validity of the PCA model is usually tested by splitting the sample set into two sets; one set for calibration and the other for internal validation (dependent test-set).

The estimation of the prediction error is in terms of the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(x_n - \hat{x}_n)^2}{D}} \tag{6}$$

where x_n and $\hat{x}_n$ are a measurement of the nth variable and its predicted (reconstructed) value, respectively. D denotes the number of degrees of freedom. The RMSE values are in the same units and scale as the reference values. Depending on how the model is used for estimating the predicted values, the following terms are used:

-RMSEC (Root Mean Square Error of Calibration)

-RMSECV (Root Mean Square Error of Cross Validation)

-RMSEP (Root Mean Square Error of Prediction)

How to divide a sample set into test set and calibration set as well as when to use cross validation and test set validation will always be related to the data set at hand and the purpose of the modeling. When choosing the number of components in the PCA model, one should try to avoid under-fitting, i.e. too few components, and over-fitting, i.e. too many components. If an insufficient number of principal components are chosen, the prediction is not reliable because useful information has been omitted. If too many components are chosen, however, more uncertainty is included in the calibration set which results in errors in prediction. When calibrating a model the RMSECV or RMSEP is usually calculated for every addition of the next component to the model. Normally, the optimal number of components is found at the first local minimum of the RMSECV or RMSEP curve.

5.5 Bootstrapping

Bootstrapping [104] is a method for estimating the distribution of a statistic that is otherwise difficult to determine because of e.g. small sample size or awkward distribution. Bootstrap methods repeatedly analyze new so-called bootstrap data sets which are created by resampling with replacement from the original data. Hence, each bootstrap data set is a random distribution of samples from the full data set. The bootstrap data set have the same number of samples as the original data set.

A wide variety of adaptations of the bootstrap have been proposed over the years, each tailored to a specific application or goal. Many of them are reviewed by Wehrens et al. (2000) [105]. In PAPER I a bootstrapping procedure was set up to empirically estimate the false positive rate, i.e. the probability of a sample being outside normal operation condition (NOC) when it is actually a NOC sample. In Figure 23 the average false positive rate is plotted against bootstrap iterations to check for convergence.

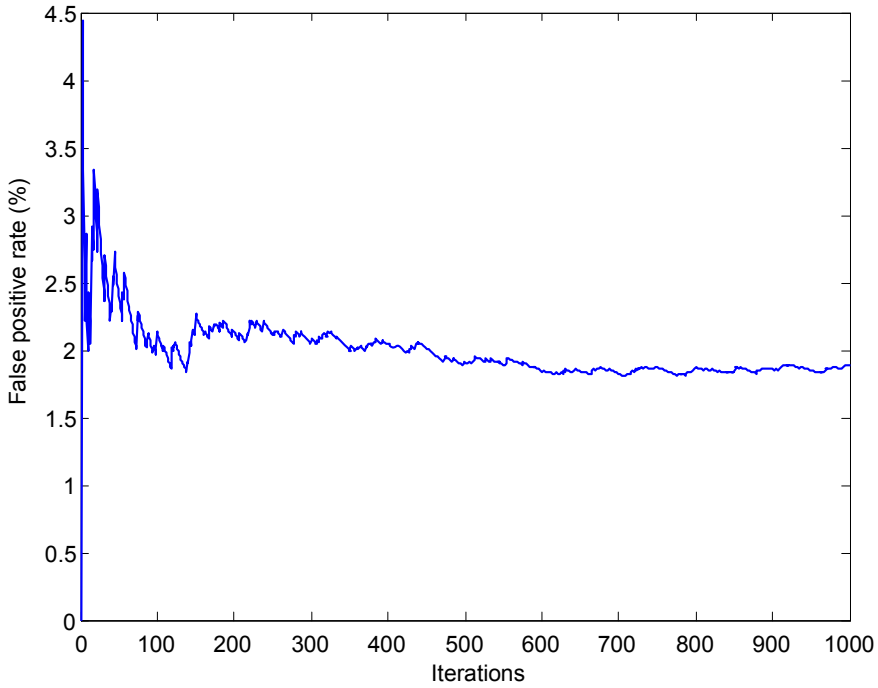

Figure 23. Average false positive rate plotted against number of bootstrap iterations. Convergence is obtained after approximately 600 bootstrap iterations [PAPER I].

The bootstrap exercise presented in Figure 23 reveals that the false positive rate estimate seems to be estimated accurately after approximately 600 iterations [PAPER I] and it seems that the rate is approximately 2%.

5.6 MSPC charts

From the PCA model two complementary multivariate monitoring statistics are commonly derived: the Hotelling T^2 (D-statistic) and the squared prediction error (Q-statistic). These two statistics can be implemented by graphical and numerical ways in two separate MSPC charts to monitor if the samples are in the accepted NOC region monitored. The sensitivity of fault detection towards changes in the NOC region depends on the historical NOC data, number of data points, preprocessing methods, and number of components included in the NOC PCA model. For both the D- and Q-statistics confidence levels can be obtained and used as control limits. There is only one limit for the D- and Q-chart instead of two for the univariate charts, and this upper control limit (UCL) can be used to detect changes from the NOC model for new independent samples. It should be noticed

that randomly induced false alarms are inherent within MSPC because of the definition of the control limits. For example, the 99% control limit states that statistically 1% of the normal operating samples will fall outside this limit and incorrectly be identified as faulty. The presence of false alarms is one of the major reasons that the process operators are skeptical of employing MSPC charts for process fault detection. Consequently, various heuristic run-rules have been suggested to signal the onset of the process fault [106]. However, for the applications described in PAPER I, II, and III, a 99.87% (~3σ) confidence level has been used as the control limit similar to the 3σ control limits used in ordinary univariate Shewart control charts. This preferably makes the control chart more reliable, despite the loss of sensitivity. If a new sample falls outside the control limit in the *D*- and/or *Q*-chart, it is characterized as a special cause and the sample is considered to deviate significantly from the NOC samples included in the PCA model.

5.6.1 D-statistic

The *D*-statistic is a measure of the variation in the PCA model, and faults detected in the *D*-chart could in chromatography mean that there is a deviation from the target value for one or more peak areas. However, the correlation structure of the peak areas remains the same. In Figure 17 this is represented as the ▲ observation showing an extreme increase in the area of both peak 1 and peak 2. In this case the correlation structure is maintained and therefore only the *D*-chart will detect this event. The *D*-statistic is described by the scores in the T^2 for principal components, introduced by Hotelling (1947) [81], and is a distance between the center of model space and the new obtained scores:

$$D_{new} = \sum_{r=1}^{R} \frac{t_{new,r}^2}{s_{t_r}^2} \tag{7}$$

where $t_{new,r}$ is the rth principal component score for the new sample, $s_{t_r}^2$ is the variance of the calibration model scores t_r of the rth component and R denote the number of principal components retained in the PCA model.

The *D*-statistic follows the *F*-distribution and the upper control limit for the *D*-statistic can be calculated according to Jackson [71]:

$$UCL_D = \frac{R(M-1)}{M-R} F_{1-\alpha}(R, M-R) \tag{8}$$

Where *M* is the number of samples, *R* is the number of principal components retained in the PCA model, and F is the F-distribution with a confidence level 1-α and (*R*,*M*-*R*) degrees of freedom.

5.6.2 Q-statistic

The *Q*-statistic is a measure of the amount of variation not captured by the PCA model, and faults detected by being extreme in the *Q*-chart are caused by events that break the correlation structure described by the model. An event related to the example in Figure 17 is the ■ observation where the area of peak 1 is decreased while the area of peak 2 is increased. Under NOC the area of peak 1 also had to increase, so in this case the new event is no longer described by the model and there will be a faulty situation only in the *Q*-chart represented in the residuals of the new sample calculated according to Jackson & Mudholkar (1979) [83]:

$$Q_{new} = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2 = \sum_{n=1}^{N} (e_n)^2 \tag{9}$$

where x_n and $\hat{x}_n$ are a new measurement of the *n*th variable and its predicted (reconstructed) value, respectively, which result in the residual e_n. *N* denotes the number of variables. Several ways to determine the UCL for the *Q*-chart is described [88,107]. Most commonly, a normal distribution to approximate a weighted chi-square distribution is used from which the UCL for the *Q*-chart can be calculated according to Jackson & Mudholkar (1979) [83]:

$$UCL_Q = \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1 - h_0}{\theta_1^2} \right) + \frac{\sqrt{z_\alpha (2\theta_2 h_0^2)}}{\theta_1} \right]^{\frac{1}{h_0}} \tag{10}$$

To understand this, **V** is defined as the covariance matrix of the residuals **E** (after performing PCA on the NOC samples), θ_1 is the trace (the sum of the elements on the main diagonal) of **V**, θ_2 the trace of $\mathbf{V}^2$, θ_3 the trace of $\mathbf{V}^3$, h_0=1-((2$\theta_1\theta_3$)/(3θ_1^2)), and z_α is the standardized normal variable with a (1-α) confidence level. Alternatively, an approximation based on the weighted chi-squared distribution ($g\chi_h^2$) can be used proposed by Box [108], with the weight g=θ_2/θ_1 and h=θ_1^2/θ_2 degrees of freedom.

In Figure 24 examples of the *D*- and *Q*-chart is presented for monitoring chromatographic profiles [PAPER II]. The chart statistics are derived from a PCA model based on forty NOC calibration samples and prediction of ten

independent NOC validation samples. The 95%, 99% and 99.73% (UCL ~3σ) confidence levels are derived from the PCA model based only on the calibration samples.

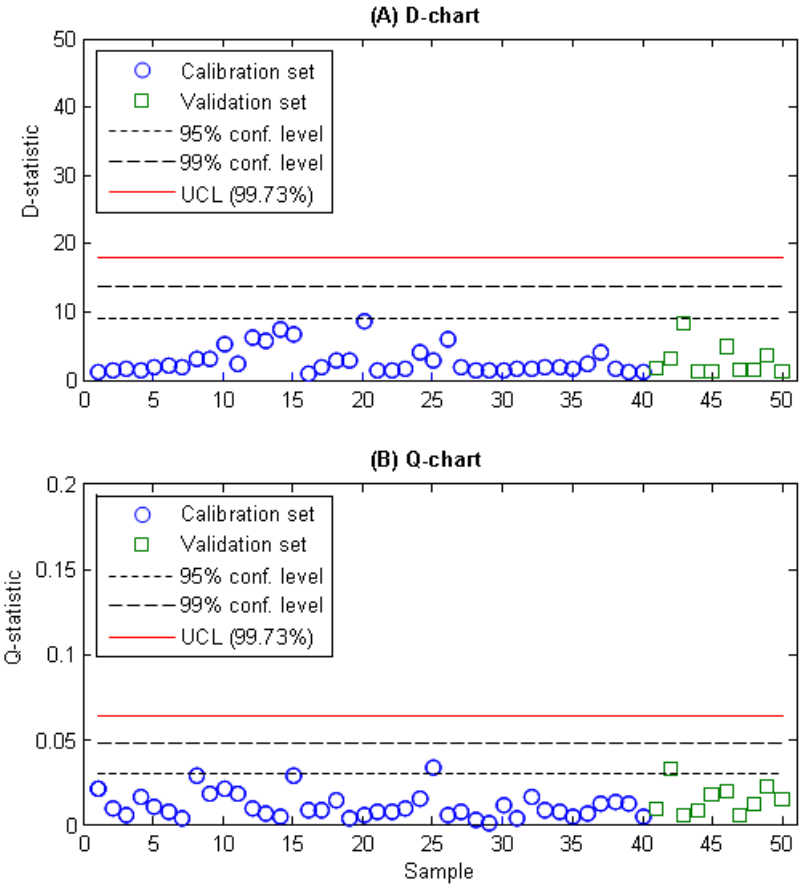

Figure 24. *D*-chart (A) and *Q*-chart (B) of calibration (circle) and validation (square) sample sets. 95%, 99% and 99.73% (~3σ) confidence levels are indicated (modified form PAPER II).

By inspection of the *D*- and *Q*-chart it can be confirmed that the PCA model based on the calibration sample set describe the common-cause variation (Figure 24). All 50 NOC samples are within the 95% confidence interval in the *D*-chart, whereas in the *Q*-chart two samples (~5%) are outside the 95% confidence interval as expected from a normal distribution point of view [PAPER II].

5.7 Contribution plots

It is not only important to detect that there is other variation in a new sample than the common-cause variation captured in the NOC samples It is also important to search for the original chromatographic cause of the fault. The *D*- and *Q*-charts do not give information on what is wrong with the detected sample, or which chromatographic signals caused the sample to be out of control. Once an MSPC chart signals an alarm, the model can be scrutinized to understand the cause of the alarm. One of the most widely used approaches is using contribution plots [109-111]. Contribution plots compute a list of each single chromatographic signal (peak area, retention time etc.) that contribute numerically to the *D*- and *Q*-statistics respectively. In this way contribution plots may reveal the group of chromatographic signals making the highest contribution to the model (*D*) or to the residuals (*Q*). If a new sample exceeds the control limit in one of the statistics or both, the contributions of each chromatographic signal to the respective statistic should be examined. In Figure 25 the residual contributions (green) of a faulty chromatogram is plotted together with the actual faulty chromatogram (red) and a NOC chromatogram (blue) (modified from PAPER II).

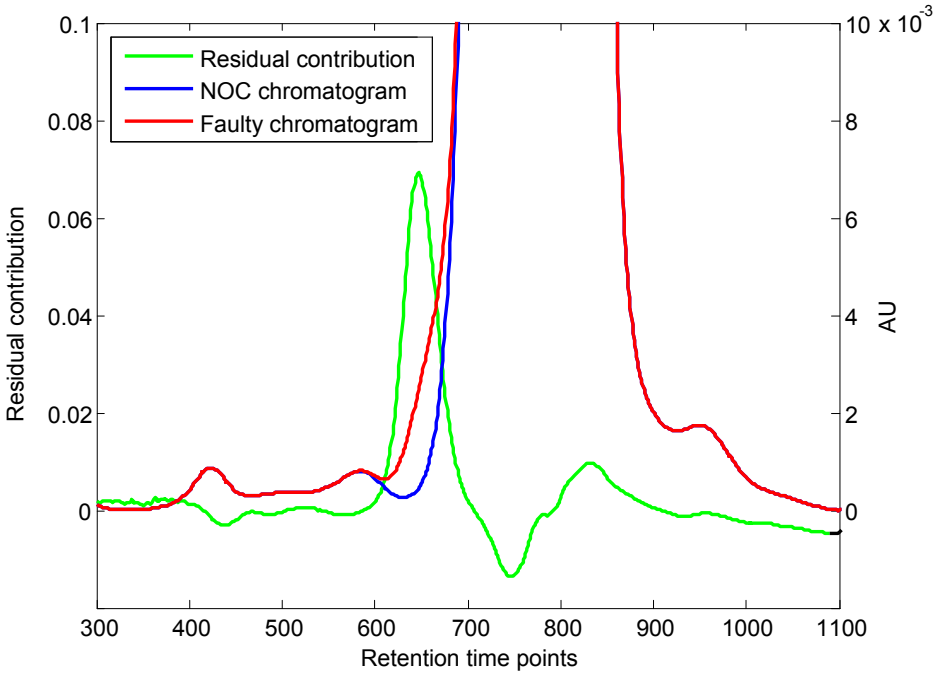

Figure 25. Plot of the faulty residual contribution (green), plotted together with a NOC (blue) and the faulty chromatogram (red) on the secondary y-axis (modified from PAPER II).

Clear indication of a new peak or a shoulder on the fronting target peak is given in Figure 25. Apparently, this variability is not described by the principal components retained in the NOC model. Another example is given in Figure 26 from PAPER I, where the integrated areas of twenty peaks are monitored. Here, a new sample is deviating in a multivariate sense, and is detected in the *Q*-chart exceeding the UCL. The chromatographic variables (peak areas) responsible for the signal in the *Q*-chart can be inspected in the residual contribution plot (Figure 26).

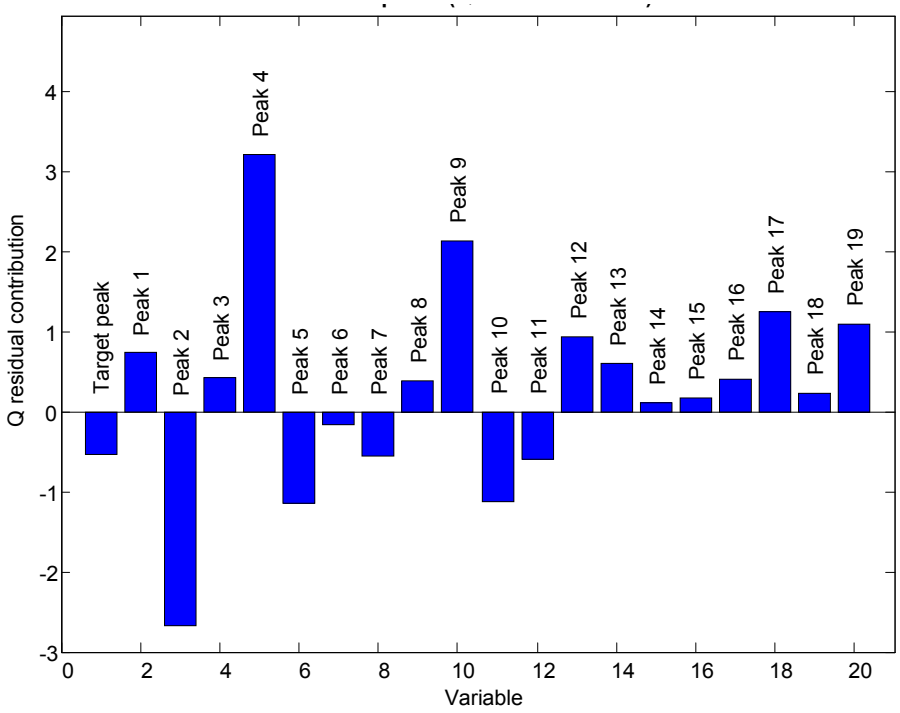

Figure 26. *Q* residual contribution plot of twenty peak areas obtained from a faulty sample exceeding the UCL in the *Q*-chart [PAPER I].

The contribution plot (Figure 26) allows us to diagnose the problem with the faulty sample, and indications of which (possibly pattern of) peaks that contribute to the deviating behavior are given.

5.8 Enhanced MSPC charts

In PAPER II the conventional MSPC *D*- and *Q*-chart is used for monitoring the chromatographic data. However, in PAPER I and PAPER III enhanced MSPC charts are developed. These are briefly described in the following subsections.

5.8.1 Comprehensive control charting (PAPER I)

The derived MSPC statistics (*D* and *Q*) may suffer from lack of sensitivity if only one or a few variables deviate from the common-cause variation in a given situation. This is simply due to the properties of PCA where a change in correlation structure is amplified over single variable changes. To comply with this phenomenon a new comprehensive control (COCO) chart procedure is developed. The COCO chart considers both univariate statistics

and multivariate statistics derived from PCA in a single plot that allows easy visualization of the combined data from a univariate and multivariate point of view. The methodology simply normalizes each control chart value (both single variables and *D*- and *Q*-values) with its respective control limit such that a value greater than one indicates deviation from normal operating conditions (NOC), whereas a value between zero and one indicates NOC. This is exemplified in Figure 27 showing three univariate control charts (one for each variable 1, 2, and 3) and the two derived multivariate control charts (*D*- and *Q*-chart), before (upper charts) and after normalization (lower charts) [PAPER I]. The control charts are based on a simulated dataset (autoscaled). The first twenty samples have been used as NOC samples, whereas the two last samples are new independent samples to be monitored. The *D*- and *Q*-charts are derived from a two component PCA model explaining approximately 80% of the common-cause variation in the first twenty NOC samples. The control limits correspond to the 3σ confidence level, and are estimated from the NOC samples.

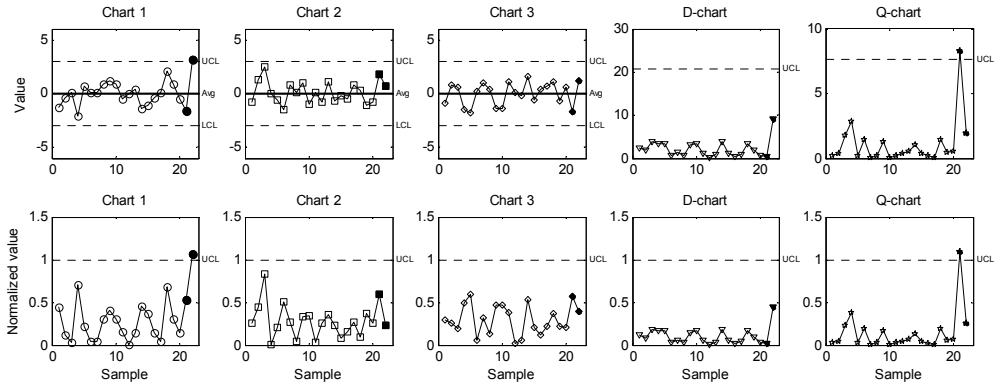

Figure 27. Simulated example of univariate- and multivariate control charts, before and after normalization with the respective control limit [PAPER I].

The two last samples (21 and 22) plotted in Figure 27 simulate two different types of special causes. Sample 21 is within common-cause variation in all three univariate charts, but clearly deviates in a multivariate sense as it exceeds the control limit in the *Q*-chart. In contrast sample 22 exceeds the control limit in the univariate Chart 1; in spite of this none of the multivariate charts detects this deviation as being faulty. Consequently, detection of these two different types of special causes would require inspection of both univariate and multivariate control charts simultaneously.

This can be an overwhelming and inefficient task and the risk of missing an out-of-control situation is obvious.

Therefore a more orderly control chart procedure is devised here. As depicted in Figure 27, Z normalized control values (Z is the number of control charts including the D- and Q-chart, here Z=5) is produced for each sample, where the largest value reflects the control chart in which the sample is most deviating. As a condensed measure across all control charts (including the D- and Q-chart) the maximum normalized value is used for COCO charting as exemplified in Figure 28.

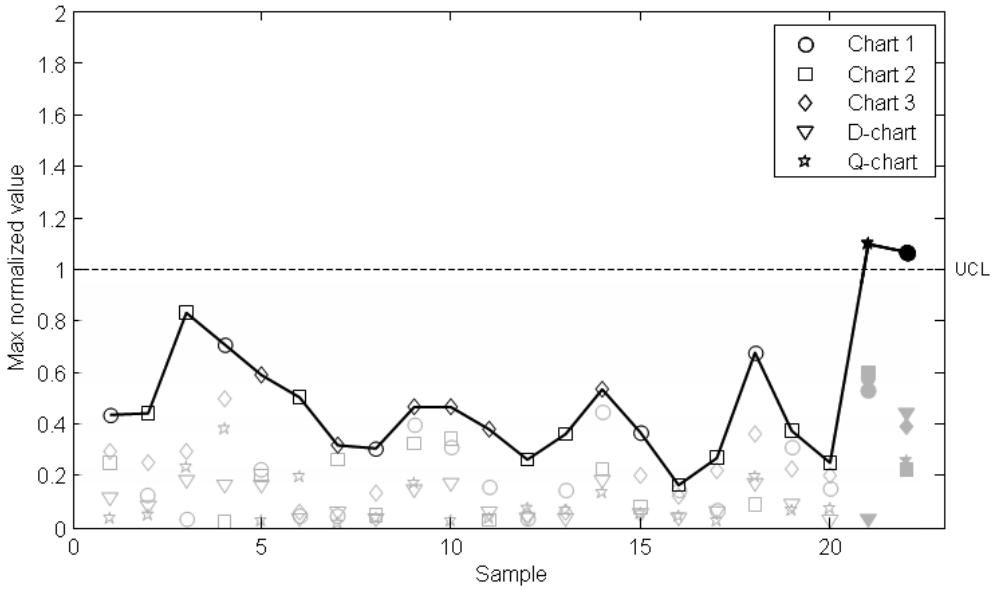

Figure 28. Simulated example of how univariate- and multivariate control charts can be condensed in one COCO chart monitoring the maximum normalized value for each sample [PAPER I].

In the COCO chart (Figure 28) the maximum fault contributions are monitored, allowing both univariate and multivariate statistics to be accounted for at the same time. As opposed to *either* using MSPC *or* using multiple SPC charts, this comprehensive control chart strategy covers the detection capabilities of both [PAPER I].

5.8.2 MSPC based on PCA combined with multiple testing (PAPER III)

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 [90,112,113]. In PAPER III MSPC is applied to LC-MS data for detection of unknown impurities. However, the huge amount of data points combined with the discrete nature of LC-MS signals (i.e. sharp signals in MS direction) makes detection of unknown impurities a case of needle-in-the-haystack expedition. That is, if a few discrete residuals are related to an unknown impurity they are simply masked when calculating the sum of squared residuals (Q), making Q a non-sensitive measure. Therefore a new method was devised to monitor the relative size of the residuals, compared to the NOC residuals, rather than just considering the absolute size of the residuals. This enhanced MSPC methodology is based on PCA in conjunction with variable wise (multiple) testing [PAPER III].

PCA and variable wise (multiple) testing offers two different dimensions to statistical data analysis. Multiple testing aims at separating the variable space into variables with a significant or non-significant change, where PCA separates data into a systematic part (D) and a non-systematic part (Q). In Figure 29 this is schematized.

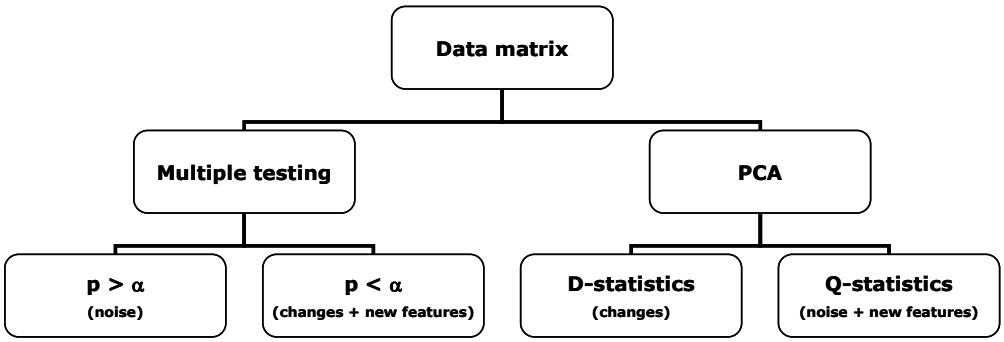

Figure 29. Shematic overview of two different data analytical approaches for extraction of information from multivariate data. p refers to test probability, α is significance level [PAPER III].

Experiments where a high number of variables are evaluated on possibly several outcomes involve testing of numerous hypotheses where handling of error rates is of crucial importance. This discipline is referred to as multiple testing. Multiple testing is widely used for biomarker discovery in proteomics, and has been applied in several analyses of LC-MS data intensities [53,114,115]. However, if multiple testing is applied directly to preprocessed LC-MS data it would result in detection of all intensity differences (i.e. both known according to normal operating conditions and

unknown features). Multiple testing applied to PCA residuals would only result in detection of unknown features, as the known features are described by the model and expressed in the *D*-statistics. In PAPER III the huge amount of data points per sample are binned into a (time, m/z) grid, where each binned value represents the sum data points within that bin. In order to detect the needle in the haystack, multiple testing is based on a simple t-test for each bin (*n*) as:

$$t_n = \frac{e_{new,n} - \bar{e}_{ref,n}}{s_n \cdot \sqrt{1 + M^{-1}}} \tag{12}$$

where

$$s_n^2 = \frac{1}{M-1} \sum_{i=1}^{M} (e_{i,n} - \bar{e}_{ref,n})^2 \tag{13}$$

and

$$\bar{e}_{ref,n} = \frac{1}{M} \sum_{i=1}^{M} e_{i,n} \tag{14}$$

where $e_{new,n}$ is the residual from the new sample for bin *n*, $\bar{e}_{ref,n}$ is the mean of the residuals from the reference samples for bin *n*. M is the number of reference samples. s_n is the standard deviation of residuals from reference samples for bin *n*.

The critical value of t is dependent on sample size. In order to remove this dependency, t is transformed to a *z*-value through a p-value:

$$P(T_{df} \leq t_n) = \Phi(z_n) \tag{15}$$

where T_{df} is the *t*-distribution with *df* degrees of freedom, *df*=*M*-1. Φ is the cumulative distribution function of the standard Gaussian distribution. This *z*-value is used as diagnostic measure for the corresponding (time, *m*/*z*) bin. The *z*-value and p-value reflects the same statistics (Equation 15) and hence the behavior of the system. In PAPER II the *Q* value was used for a new sample as a measure for detecting subtle differences in the chromatographic pattern. The methodology devised in PAPER III produces not one but *K* significance tests where *K* is the number of bins. These are expressed as a list of *z*-values; $z_1, z_2, .., z_K$,. The largest values of $z_1, z_2, .., z_K$ reflect the variables where the new sample is most deviating. Impurities are in excess and hence only large positive *z*-values are of interest. The present method proposes use of the *maximum* *z*-value across all bins as a measure in control chart monitoring [PAPER III].

6 Conclusions and perspectives

This thesis has focused on solutions providing more comprehensive monitoring capabilities of analytical chromatographic data in the pharmaceutical industry. The research presented in this thesis has demonstrated the unique potentials of assessing chromatographic data using novel multivariate statistical tools. These tools utilize the available information contained in multiple measured chromatographic signals simultaneously in an objective (numerical) and statistically reliable way.

Methods and algorithms have been developed to automate and optimize the many aspects present, when setting up an industrial reliable monitoring scheme. This includes:

- Collection of data from commercial chromatographic instruments to numerical software (MATLAB)
- Application of necessary preprocessing steps to generate 'clean' data
- Multivariate statistical modeling based on PCA and multiple testing
- Comprehensive control chart monitoring and detection
- Interpretable visualizations providing diagnostic information on deviating chromatographic data

These new and useful tools have been presented, explained and visualized on actual pharmaceutical analytical chromatographic data and published in three scientific papers.

In PAPER I it was demonstrated how multivariate statistical process control (MSPC) based on principal component analysis (PCA) is a much more powerful tool for detecting variations, due to special causes than conventional single variable statistical process control (SPC). Furthermore, the PCA based SPC simplifies monitoring as it limits the number of control charts to typically two charts rather than one for each signal. However, the derived MSPC statistics may suffer from lack of sensitivity if only one or a few variables deviate in a given situation. A new comprehensive control (COCO) chart procedure was developed that considers both univariate statistics and multivariate statistics derived from PCA in a single plot that allows easy visualization of the combined data from a univariate and multivariate point of view. The method was exemplified using integrated

areas of twenty chromatographic peaks obtained for purity analysis of a biopharmaceutical in-process sample. The new control chart procedure may serve as a powerful supplement to the current univariate chromatographic data approach used in the industry.

PAPER II proposes a PCA-based MSPC approach for monitoring subtle changes in the chromatographic profile, providing clear diagnostics of subtly deviating chromatograms due to new impurities co-eluting with the target compound (usually present in excess compared to any impurity). Different chromatographic data preprocessing methods such as time alignment, baseline correction and scaling were applied to historical chromatograms from a biopharmaceutical in-process analysis to correct for non-relevant analytical variation, since it largely influences the outcome of the monitoring. The procedure can be implemented and operated as the chromatographic analysis runs, and support the current practiced visual inspection of chromatograms. In this way an automated and timely tool for continuous quality verification of the chromatographic data is conducted in an objective and statistically reliable way.

PAPER III describes how LC-MS adds a new selective dimension to the chromatographic separation in order to increase confidence that all impurities are detected. The study demonstrates how the relevant chemical information can be extracted from the huge amount of data generated with LC-MS analysis. This is particularly helpful when the presence of unknown impurities is investigated. In PAPER III a new tool, based on PCA combined with multiple testing, was developed to adapt MSPC based monitoring to the nature of LC-MS data. The tool was applied to LC-MS data from in-process analysis of industrial insulin intermediate samples. The study demonstrated, how low spike-levels (0.05%) of a structurally related compounds co-eluting with the target compound was detected by the tool and further how clear diagnostics of the co-eluting compound was provided. This tool makes a fully automated monitoring of LC-MS data possible, where only relevant areas in the LC-MS data are highlighted for further interpretation.

By developing and demonstrating ways to improve assessment of chromatographic data, this thesis is a step in the direction of better utilization of available information present in the data-rich pharmaceutical industry. The applications described in PAPER I-III may all serve as complementary and equally important approaches for assessment of various types of chromatographic data. This will be a step toward effectiveness and

robustness, and consequently enhance the overall chromatographic analysis significantly. Of course, these new MSPC tools are not just *plug and play,* but may need increased allocation of resources, compared to common SPC tools, for development, implementation, and maintenance. This will require close interaction between analytical chemists, process operators, and experts of advanced data analytical techniques. However, regulatory requirements on documentation and validation of non-commercial MSPC systems can be quite extensive, and a rather laborious task to fulfill. Therefore, the pharmaceutical industry should push for improved validated commercial instrumentation software where these MSPC tools are integrated.

Future pharmaceutical process analysis will continuously develop towards handling more complex samples at increasingly higher speed. This will require even more advanced analytical instruments. Consequently, the amount and complexity of the acquired analytical data will increase, and the role of multivariate statistical tools may become a necessity for optimal use of such new sophisticated analytical instrumentation. This PhD thesis encourages to speed up the inclusion of more advanced data analytical tools in validated commercial instrumentation software or database management systems. In this way advanced tools such as MSPC will become more user-friendly, familiar to a broader range of end-users, and ultimately facilitate optimal utilization of available information.

7 References

[1] R. L. Garnick, N. J. Solli, and P. A. Papa, The Role of Quality-Control in Biotechnology - An Analytical Perspective, Analytical Chemistry, 60 (1988) 2546-2557.

[2] Council of Europe, European Pharmacopoeia, 2007.

[3] United States Pharmacopoeia (USP), US 26-NF 21, 2003.

[4] U.S.Food and Drug Administration (FDA) - Department of Health and Human Services. Pharmaceutical cGMPs for the 21st Century - A Risk-based Approach, Final Report. 2004.

[5] International Conference on Harmonisation (ICH) - Guidance for industry: Q8(R2) Pharmaceutical Development. Q8(R2) Pharmaceutical Development. 2009.

[6] International Conference on Harmonisation (ICH) - Guidance for industry: Q9 Quality Risk Management. Q9 Quality Risk Management. 2005.

[7] P. F. Gavin and B. A. Olsen, A quality by design approach to impurity method development for atomoxetine hydrochloride (LY139603), Journal of Pharmaceutical and Biomedical Analysis, 46 (2008) 431-441.

[8] P. McKenzie, S. Kiang, J. Tom, A. E. Rubin, and M. Futran, Can pharmaceutical process development become high tech?, Aiche Journal, 52 (2006) 3990-3994.

[9] L. X. Yu, Pharmaceutical quality by design: Product and process development, understanding, and control, Pharmaceutical Research, 25 (2008) 781-791.

[10] D. Jacobson-Kram and T. McGovern, Toxicological overview of impurities in pharmaceutical products, Advanced Drug Delivery Reviews, 59 (2007) 38-42.

[11] R. N. Rao and V. Nagaraju, An overview of the recent trends in development of HPLC methods for determination of impurities in drugs, Journal of Pharmaceutical and Biomedical Analysis, 33 (2003) 335-377.

[12] International Conference on Harmonization (ICH) - Guidance for Industry: Q3A(R2) Impurities in New Drug Substances. Q3A(R2) Impurities in New Drug Substances. 2006.

[13] M. D. Argentine, P. K. Owens, and B. A. Olsen, Strategies for the investigation and control of process-related impurities in drug substances, Advanced Drug Delivery Reviews, 59 (2007) 12-28.

[14] K. M. Alsante, A. Ando, R. Brown, J. Ensing, T. D. Hatajik, W. Kong, and Y. Tsuda, The role of degradant profiling in active pharmaceutical ingredients and drug products, Advanced Drug Delivery Reviews, 59 (2007) 29-37.

[15] F. H. Qiu and D. L. Norwood, Identification of pharmaceutical impurities, Journal of Liquid Chromatography & Related Technologies, 30 (2007) 877-935.

[16] International Conference on Harmonization (ICH). Q2(R1) Validation of Analytical Procedures. 1994.

[17] D. C. Montgomery, Introduction to Statistical Quality Control, John Wiley and Sons, Hoboken, NJ 2005.

[18] US Food and Drug Administration. PAT guidance for industry - a framework for innovative pharmaceuticaldevelopment,manufacturing and quality assurance. 2004. Rockville, MD.

[19] H. J. Ramaker, E. N. M. van Sprang, S. P. Gurden, J. A. Westerhuis, and A. K. Smilde, Improved monitoring of batch processes by incorporating external information, Journal of Process Control, 12 (2002) 569-576.

[20] H. J. Ramaker, E. N. M. van Sprang, J. A. Westerhuis, and A. K. Smilde, The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring, Chemometrics and Intelligent Laboratory Systems, 73 (2004) 181-187.

[21] A. Ferrer, Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process, Quality Engineering, 19 (2007) 311-325.

[22] J. M. Miller, Chromatography: Concepts and Contrasts, John Wiley & Sons, Inc., New York 2005.

[23] E. d. Hoffmann and V. Stroobant, Mass Spectrometry: Principles and Applications, John Wiley & Sons, Ltd, 2007.

[24] R. E. Ardrey, Liquid chromatography - mass spectrometry: An introduction, John Wiley & Sons, 2003.

[25] J. T. Watson and O. D. Sparkman, Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation, John Wiley & Sons, Ltd, 2007.

[26] S. Kromidas and H.-J. Kuss, Quantification in LC and GC: A Practical Guide to Good Chromatographic Data, WILEY-VCH, 2009.

[27] M. Tswett, Adsorptionanalyse und chromatographische Methode - Anwendung auf die Chemie des Chlorophylls (Adsorption analysis and chromatographic method - Application to the chemistry of chlorophyll), Berichte der Deutschen botanischen Gesellschaft, 24 (1906) 384-393.

[28] L. S. Ettre, M.S. Tswett and the invention of chromatography, Lc Gc North America, 21 (2003) 458-467.

[29] R. J. P. Martin and R. L. M. Synge, A new form of chromatogram employing two liquid phases. A theory of chromatography 2. Application to the micro-determination of the higher monoamino-acidsin proteins, Biochemical Journal, 35 (1941) 1358-1368.

[30] R. S. Alm, Gradient Elution Analysis .2. Oligosaccharides, Acta Chemica Scandinavica, 6 (1952) 1186-1193.

[31] L. S. Ettre, Gas chromatography - Past, present, and future, Lc Gc North America, 19 (2001) 120-123.

[32] I. Ali, H. Y. Aboul-Enein, and J. Cazes, A Journey from Mikhail Tswett to Nano-Liquid Chromatography, Journal of Liquid Chromatography & Related Technologies, 33 (2010) 645-653.

[33] V. L. Talroze, G. V. Karpov, I. G. Gorodets, and V. E. Skurat, Capillary System for Introduction of Liquid Mixtures Into An Analytical Mass-Spectrometer, Russian Journal of Physical Chemistry, USSR, 42 (1968) 1658.

[34] C. R. Blakley and M. L. Vestal, Thermospray Interface for Liquid-Chromatography Mass-Spectrometry, Analytical Chemistry, 55 (1983) 750-754.

[35] R. M. Caprioli, T. Fan, and J. S. Cottrell, Continuous-Flow Sample Probe for Fast-Atom-Bombardment Mass-Spectrometry, Analytical Chemistry, 58 (1986) 2949-2954.

[36] C. M. Whitehouse, R. N. Dreyer, M. Yamashita, and J. B. Fenn, Electrospray Interface for Liquid Chromatographs and Mass Spectrometers, Analytical Chemistry, 57 (1985) 675-679.

[37] E. C. Horning, D. I. Carroll, I. Dzidic, K. D. Haegele, M. G. Horning, and R. N. Stillwel, Liquid Chromatograph Mass Spectrometer-Computer Analytical Systems - Continuous-Flow System Based on Atmospheric-Pressure Ionization Mass-Spectrometry, Journal of Chromatography, 99 (1974) 13-21.

[38] S. F. Wong, C. K. Meng, and J. B. Fenn, Multiple Charging in Electrospray Ionization of Poly(Ethylene Glycols), Journal of Physical Chemistry, 92 (1988) 546-550.

[39] C. K. Lim and G. Lord, Current developments in LC-MS for pharmaceutical analysis, Biological & Pharmaceutical Bulletin, 25 (2002) 547-557.

[40] B. Steffen, K. P. Muller, M. Komenda, R. Koppmann, and A. Schaub, A new mathematical procedure to evaluate peaks in complex chromatograms, Journal of Chromatography A, 1071 (2005) 239-246.

[41] N. P. V. Nielsen, J. Smedsgaard, and J. C. Frisvad, Full second-order chromatographic/spectrometric data matrices for automated sample identification and component analysis by non-data-reducing image analysis, Analytical Chemistry, 71 (1999) 727-735.

[42] M. Daszykowski and B. Walczak, Use and abuse of chemometrics in chromatography, TrAC Trends in Analytical Chemistry, 25 (2006) 1081-1096.

[43] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems, 38 (1997) 149-171.

[44] Peak detection, in: A. Felinger (Ed.), Data Handling in Science and Technology. Data Analysis and Signal Processing in Chromatography, Elsevier, 1998, 183-190.

[45] M. K. L. Bicking, Integration errors in chromatographic analysis, part II: Large peak size ratios, Lc Gc North America, 24 (2006) 604-615.

[46] R. G. Harrison, P. Todd, S. R. Rudge, and D. P. Petrides, Bioseparations Science and Engineering, Oxford University Press, New York 2003.

[47] L. S. Ettre, Nomenclature for Chromatography, Pure and Applied Chemistry, 65 (1993) 819-872.

[48] K. Wiberg, M. Andersson, A. Hagman, and S. P. Jacobsson, Peak purity determination with principal component analysis of high-performance liquid chromatography-diode array detection data, Journal of Chromatography A, 1029 (2004) 13-20.

[49] J. D. Wilson and C. A. J. Mcinnes, Elimination of Errors Due to Baseline Drift in Measurement of Peak Areas in Gas Chromatography, Journal of Chromatography, 19 (1965) 486.

[50] P. H. C. Eilers, Parametric time warping, Analytical Chemistry, 76 (2004) 404-411.

[51] F. Gan, G. H. Ruan, and J. Y. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemometrics and Intelligent Laboratory Systems, 82 (2006) 59-65.

[52] van den Berg, F. Baseline_spline: Determines splines-based baseline by gradually eliminating points (www.models.life.ku.dk). 2008 (unpublished work).

[53] J. Listgarten and A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry, Molecular & Cellular Proteomics, 4 (2005) 419-434.

[54] M. E. Parrish, B. W. Good, F. S. Hsu, F. W. Hatch, D. M. Ennis, D. R. Douglas, J. H. Shelton, D. C. Watson, and C. N. Reilley, Computer-Enhanced High-Resolution Gas-Chromatography for the Discriminative Analysis of Tobacco-Smoke, Analytical Chemistry, 53 (1981) 826-831.

[55] J. A. Pino, J. E. Mcmurry, P. C. Jurs, B. K. Lavine, and A. M. Harper, Application of Pyrolysis-Gas Chromatography Pattern-Recognition to the Detection of Cystic-Fibrosis Heterozygotes, Analytical Chemistry, 57 (1985) 295-302.

[56] M. D. Hamalainen, Y. Z. Liang, O. M. Kvalheim, and R. Andersson, Deconvolution in One-Dimensional Chromatography by Heuristic Evolving Latent Projections of Whole Profiles Retention Time Shifted by Simplex Optimization of Cross-Correlation Between Target Peaks, Analytica Chimica Acta, 271 (1993) 101-114.

[57] N. P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A, 805 (1998) 17-35.

[58] G. Malmquist and R. Danielsson, Alignment of Chromatographic Profiles for Principal Component Analysis - A Prerequisite for Fingerprinting Methods, Journal of Chromatography A, 687 (1994) 71-88.

[59] D. Bylund, R. Danielsson, G. Malmquist, and K. E. Markides, Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data, Journal of Chromatography A, 961 (2002) 237-244.

[60] K. J. Johnson, B. W. Wright, K. H. Jarman, and R. E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, Journal of Chromatography A, 996 (2003) 141-155.

[61] G. Tomasi, F. van den Berg, and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, Journal of Chemometrics, 18 (2004) 231-241.

[62] T. Skov, F. van den Berg, G. Tomasi, and R. Bro, Automated alignment of chromatographic data, Journal of Chemometrics, 20 (2006) 484-497.

[63] F. van den Berg, G. Tomasi, N. Viereck, and S. B. Engelsen, Magnetic Resonance in Food Science: The Multivariate Challenge, in: P. S. Belton and H. J. Jakobsen (Eds.), Cambridge, 2005, p. 131.

[64] F. Savorani, G. Tomasi, and S. B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, Journal of Magnetic Resonance, 202 (2010) 190-202.

[65] V. Pravdova, B. Walczak, and D. L. Massart, A comparison of two algorithms for warping of analytical signals, Analytica Chimica Acta, 456 (2002) 77-92.

[66] M. Vandenbogaert, S. Li-Thiao-Te, H. M. Kaltenbach, R. X. Zhang, T. Aittokallio, and B. Schwikowski, Alignment of LC-MS images, with applications to biomarker discovery and protein identification, Proteomics, 8 (2008) 650-672.

[67] J. H. Christensen, G. Tomasi, and A. B. Hansen, Chemical fingerprinting of petroleum biomarkers using time warping and PCA, Environmental Science & Technology, 39 (2005) 255-260.

[68] M. Daszykowski and B. Walczak, Target selection for alignment of chromatographic signals obtained using monochannel detectors, Journal of Chromatography A, 1176 (2007) 1-11.

[69] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, Bmc Genomics, 7 (2006).

[70] R. Bro and A. K. Smilde, Centering and scaling in component analysis, Journal of Chemometrics, 17 (2003) 16-33.

[71] J. E. Jackson, A user's guide to principal components, John Wiley and Sons, New York 1991.

[72] T. Kourti, J. Lee, and J. F. MacGregor, Experiences with industrial applications of projection methods for multivariate statistical process control, Computers & Chemical Engineering, 20 (1996) S745-S750.

[73] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24 (1933) 417-441.

[74] R. Bro, C. A. Andersson, and H. A. L. Kiers, PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics, 13 (1999) 295-309.

[75] S. Wold, M. Sjostrom, and L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, 58 (2001) 109-130.

[76] W. A. Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand Reinhold, Princeton, NJ 1931.

[77] E. S. Page, Continuous Inspection Schemes, Biometrika, 41 (1954) 100-115.

[78] R. H. Woodward and P. L. Goldsmith, Cumulative Sum Techniques, Oliver and Boyd, Edinburgh, 1964, pp. 1-65.

[79] J. S. Hunter, The Exponentially Weighted Moving Average, Journal of Quality Technology, 18 (1986) 203-210.

[80] A. Nijhuis, S. de Jong, and B. G. M. Vandeginste, The application of multivariate quality control in gas chromatography, Chemometrics and Intelligent Laboratory Systems, 47 (1999) 107-125.

[81] H. Hotelling, Multivariate Quality Control. In: Techniques of Statistical Analysis, in: C. Eisenhart, M. W. Hastey, and W. A. Wallis (Eds.), McGraw-Hill, New York, 1947, pp. 111-184.

[82] J. E. Jackson, Quality Control Methods for Several Related Variables, Technometrics, 1 (1959) 359-377.

[83] J. E. Jackson and G. S. Mudholkar, Control Procedures for Residuals Associated with Principal Component Analysis, Technometrics, 21 (1979) 341-349.

[84] J. V. Kresta, J. F. MacGregor, and T. E. Marlin, Multivariate Statistical Monitoring of Process Operating Performance, Canadian Journal of Chemical Engineering, 69 (1991) 35-47.

[85] B. M. Wise, N. L. Ricker, D. F. Veltkamp, and B. R. Kowalski, Theoretical basis for the use of principal component models for monitoring multivariate processes, Process Control and Quality, 1 (1990) 41-51.

[86] T. Kourti and J. F. MacGregor, Multivariate SPC methods for process and product monitoring, Journal of Quality Technology, 28 (1996) 409-428.

[87] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, International Journal of Adaptive Control and Signal Processing, 19 (2005) 213-246.

[88] S. J. Qin, Statistical process monitoring: basics and beyond, Journal of Chemometrics, 17 (2003) 480-502.

[89] S. Bersimis, S. Psarakis, and J. Panaretos, Multivariate statistical process control charts: An overview, Quality and Reliability Engineering International, 23 (2007) 517-543.

[90] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, Standardized Q-statistic for improved sensitivity in the monitoring of residuals in MSPC, Journal of Chemometrics, 14 (2000) 335-349.

[91] A. Nijhuis, S. deJong, and B. G. M. Vandeginste, Multivariate statistical process control in chromatography, Chemometrics and Intelligent Laboratory Systems, 38 (1997) 51-62.

[92] S. Kittiwachana, D. L. S. Ferreira, L. A. Fido, D. R. Thompson, R. E. A. Escott, and R. G. Brereton, Dynamic analysis of on-line high-performance liquid chromatography for multivariate statistical process control, Journal of Chromatography A, 1213 (2008) 130-144.

[93] L. F. Zhu, R. G. Brereton, D. R. Thompson, P. L. Hopkins, and R. E. A. Escott, On-line HPLC combined with multivariate statistical process control for the monitoring of reactions, Analytica Chimica Acta, 584 (2007) 370-378.

[94] M. Fransson, A. Sparen, B. Lagerholm, and L. Karlsson, On-line process control of liquid chromatography, Analytical Chemistry, 73 (2001) 1502-1508.

[95] L. Schweitz, M. Fransson, L. Karlsson, A. Torstensson, and E. Johansson, On-line process control of gradient elution liquid chromatography, Analytical Chemistry, 76 (2004) 4875-4880.

[96] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2 (1901) 559-572.

[97] B. R. Kowalski, Chemometrics, Analytical Chemistry, 52 (1980) 112-122.

[98] S. Wold, K. Esbensen, and P. Geladi, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems, 2 (1987) 37-52.

[99] R. G. Brereton, Chemometrics in Analytical-Chemistry - A Review, Analyst, 112 (1987) 1635-1657.

[100] K. Esbensen, Multivariate Data Analysis - In Practice, Camo Process AS, Oslo, Norway 2002.

[101] B. M. Wise, N. B. Gallagher, S. W. Butler, D. D. White, and G. G. Barna, A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, Journal of Chemometrics, 13 (1999) 379-396.

[102] S. Wold, Cross-Validatory Estimation of Number of Components in Factor and Principal Components Models, Technometrics, 20 (1978) 397-405.

[103] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers, Cross-validation of component models: A critical look at current methods, Analytical and Bioanalytical Chemistry, 390 (2008) 1241-1251.

[104] B. Efron, Nonparametric Estimates of Standard Error - the Jackknife, the Bootstrap and Other Methods, Biometrika, 68 (1981) 589-599.

[105] R. Wehrens, H. Putter, and L. M. C. Buydens, The bootstrap: a tutorial, Chemometrics and Intelligent Laboratory Systems, 54 (2000) 35-52.

[106] E. B. Martin and A. J. Morris, Non-parametric confidence bounds for process performance monitoring charts, Journal of Process Control, 6 (1996) 349-358.

[107] P. Nomikos and J. F. MacGregor, Multivariate Spc Charts for Monitoring Batch Processes, Technometrics, 37 (1995) 41-59.

[108] G. E. P. Box, Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems .2. Effects of Inequality of Variance and of Correlation Between Errors in the 2-Way Classification, Annals of Mathematical Statistics, 25 (1954) 484-498.

[109] P. Miller, R. E. Swanson, and C. E. Heckler, Contribution plots: the missing link in multivariate quality control, *37th Annual Fall Conference ASQC*, 1993.

[110] J. F. MacGregor and T. Kourti, Statistical Process-Control of Multivariate Processes, Control Engineering Practice, 3 (1995) 403-414.

[111] J. A. Westerhuis, S. P. Gurden, and A. K. Smilde, Generalized contribution plots in multivariate statistical process monitoring, Chemometrics and Intelligent Laboratory Systems, 51 (2000) 95-114.

[112] P. Ralston, G. Depuy, and J. H. Graham, Computer-based monitoring and fault diagnosis: a chemical process case study, Isa Transactions, 40 (2001) 85-98.

[113] P. Ralston, G. Depuy, and J. H. Graham, Graphical enhancement to support PCA-based process monitoring and fault diagnosis, Isa Transactions, 43 (2004) 639-653.

[114] M. C. Wiener, J. R. Sachs, E. G. Deyanova, and N. A. Yates, Differential mass spectrometry: A label-free LC-MS method for finding significant differences in complex peptide and protein mixtures, Analytical Chemistry, 76 (2004) 6085-6096.

[115] J. Listgarten, R. M. Neal, S. T. Roweis, P. Wong, and A. Emili, Difference detection in LC-MS data for protein biomarker discovery, Bioinformatics, 23 (2007) E198-E204.

PAPER I

Laursen K., Rasmussen M.A., Bro R.

Comprehensive control charting applied to chromatography

Chemometrics and Intelligent Laboratory Systems

107 (2011) 215-225

Chemometrics and Intelligent Laboratory Systems 107 (2011) 215–225

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Comprehensive control charting applied to chromatography

Kristoffer Laursen [a,b,*], Morten A. Rasmussen [a], Rasmus Bro [a]

[a] Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30 DK-1958, Frederiksberg C Denmark
[b] Novo Nordisk A/S, Novo Allé 6 DK-2880, Bagsværd Denmark

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form 28 March 2011
Accepted 1 April 2011
Available online 8 April 2011

Keywords:
Multivariate statistical process control (MSPC)
Principal component analysis (PCA)
Bootstrapping
False positive rate
Analytical chromatography

ABSTRACT

Multivariate statistical process control (MSPC) based for example on principal component analysis (PCA) can make use of the information contained in multiple measured signals simultaneously. This can be much more powerful in detecting variations due to special causes than conventional single variable statistical process control (SPC). Furthermore, the PCA based SPC simplifies monitoring as it limits the number of control charts to typically two charts rather than one for each signal. However, the derived MSPC statistics may suffer from lack of sensitivity if only one or a few variables deviate in a given situation. In this paper we develop a new comprehensive control (COCO) chart procedure that considers both univariate statistics and multivariate statistics derived from PCA in a single plot that allows easy visualization of the combined data from a univariate and multivariate point of view. The method is exemplified using twenty analytical chromatographic peak areas obtained for purity analysis of a biopharmaceutical drug substance. The new control chart procedure detected two different types of faulty events in this study.

1. Introduction

Typical purity analysis based on high performance liquid chromatography (HPLC) in biopharmaceutical processes usually deals with a number of well known peaks of the target compound and related impurity compounds. Commonly the concentration of each compound of interest is investigated with a separate control chart. A univariate statistical process control (SPC) chart can e.g. be of the Shewart type [1], which is a simple plot of the compound vs. time, sample or batch. Such a chart usually consists of two control limits (target or mean value $\pm 3\sigma$) which indicate the range of acceptable variation of the compound. Applying univariate SPC charts to an in-process analysis containing several impurity compounds will force the practitioner to inspect a large number of control charts. The risk of making mistakes is higher when several control charts are to be checked [2]. When special events occur in a process they affect not only the magnitude of the compounds but also their relationship to each other. These events are often difficult to detect by charting one compound at a time because the correlations between the compounds is not directly affected in the individual charts.

The major benefit of Multivariate SPC (MSPC) compared to univariate SPC is that the correlation between the original variables is considered, which decreases the risk of missing an out-of-control situation due to a change in the *pattern* of variation. In MSPC the information contained within all of the variables is reduced down to a few common dimensions through the application projection methods such as principal component analysis (PCA) [3]. In the chromatographic discipline, MSPC based on PCA has also found its use [4–8]. Using the information contained in all the measured signals simultaneously, MSPC charts have shown to be much more powerful in detecting special causes than conventional single variable SPC charts [9,10]. Special causes detected in the derived MSPC charts can either be due to deviation from *common-cause variation* (detected in Q-statistic) and/or in the *magnitude* of the common cause variation (detected in D-statistic). However, these derived statistics may suffer from lack of sensitivity if only one or a few variables deviate from the common-cause variation. This is simply due to the properties of PCA where a change in correlation structure is amplified over single variable changes. To comply with this phenomenon there is a need for a comprehensive monitoring tool, that take all kinds of special causes into account. This study devises a single overall control chart for comprehensive monitoring of individual levels as well as common cause level. The method is applied to twenty defined peaks in analytical chromatography obtained for purity analysis of a biopharmaceutical drug substance.

2. Theory and methods

2.1. SPC vs. MSPC

Most statistical process control (SPC) approaches are based upon the control charting of a small number of variables, and examining them one at a time (univariately). This is inappropriate for many process applications where several variables are generated and where

* Corresponding author at: Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30 DK-1958, Frederiksberg C, Denmark.
E-mail address: krfl@novonordisk.com (K. Laursen).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.04.002

the variations in these variables are correlated. The practitioner can not really study more than two or three charts to maintain overview of the process. Furthermore, the univariate control charts do not explicitly account for the correlation structure in the data. This is exemplified in Fig. 1.

In Fig. 1, a two-dimensional data set composed of the areas of two chromatographic peaks is presented in both a univariate- and a multivariate manner (one plotted vs. the other in a scatter plot to reveal correlations). The ellipse in the scatter plot represents the common correlation structure in the data. In order to compare the univariate statistical approach with the multivariate approach, the univariate control charts of peak 1 and peak 2 are given. All nine ● samples are describing common-cause variation both in a univariate- and a multivariate sense. The ▲ sample does not deviate from the correlation structure but is clearly an extreme both in a univariate and a multivariate sense. The ■ sample seems to be within common-cause variation in a univariate sense, but clearly deviates in a multivariate sense. This is caused by the fact that the ■ observation departs from the correlation structure in the data. The univariate charts are clearly missing a faulty situation as a consequence of the correlation structure in the data which is not accounted for in the univariate approach. The principle of multivariate control charts is of course of more interest when one has to deal with a higher dimensional data set, for instance several chromatographic peaks.

2.2. MSPC based on PCA

The basis of MSPC is to collect a set of historical data when the process is running under normal operating condition (NOC). Then the multivariate statistical technique PCA is applied to the historical data to model and extract the correlation structure of several correlated variables. The data matrix $\mathbf{X}$ (with M rows of samples and N columns of variables) is decomposed into R ($R \leq \min(M,N)$) principal components $\mathbf{TP}^{\mathrm{T}}$ and a residual part $\mathbf{E}$ ($M \times N$):

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \quad (1)$$

where $\mathbf{T}$ ($M \times R$) is the score matrix and $\mathbf{P}$ ($N \times R$) is the loading matrix, with R components. $\hat{\mathbf{X}}$ is the PCA approximation of the original data.

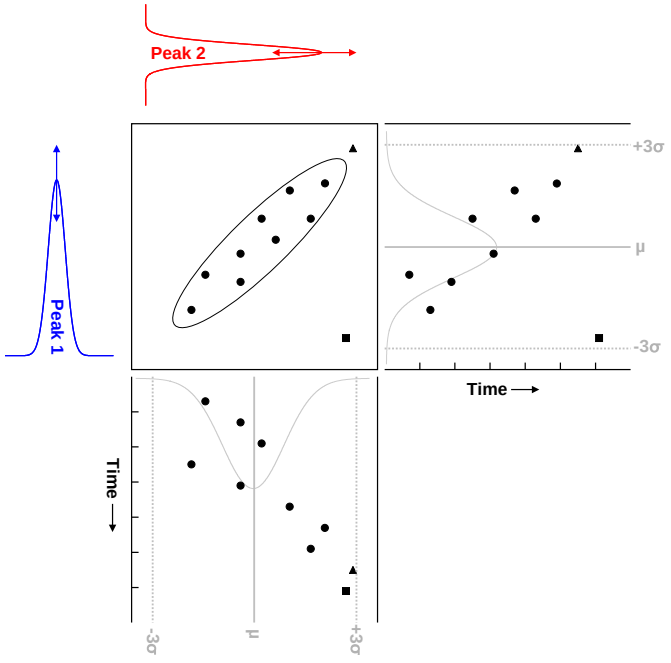

Fig. 1. Outline of different situations in univariate control chart monitoring (modified from Nijhuis et al., 1999 [5]).

The number of significant principal components can be determined by cross-validation [11]. In this way the dimensionality of the data matrix is reduced while capturing the underlying relationship between the variables. From the PCA model two complementary multivariate monitoring statistics are produced, the D-statistic and the Q-statistic. These two statistics can be monitored in separate MSPC charts.

Faults detected in the D-chart could in chromatography mean that there is a deviation from the target value for one or more peak areas. However, the correlation structure of the peak areas (the peak 'pattern') remains the same. In Fig. 1 this is represented as the ▲ sample showing an extreme increase in the area of both peak 1 and peak 2. In this case the correlation structure is maintained and therefore only the D-chart will detect this event. The D-statistic is described by the scores in the T^2 for principal components, introduced by Hotelling (1947) [12], and is a Mahalanobis distance between the center of model space and the new obtained scores:

$$D_{new} = \sum_{r=1}^{R} \frac{t_{new,r}^2}{s_{t_r}^2} \quad (2)$$

where $t_{new,r}$ is the rth principal component score for the new sample, $s_{t_r}^2$ is the variance of the model scores t_r of the rth component and R denote the number of principal components retained in the PCA model. The D-statistic follows the scaled F-distribution and the upper control limit (UCL) for the D-statistic can be calculated according to Jackson [13]:

$$UCL_D = \frac{R(M-1)}{M-R} F_{1-\alpha}(R, M-R) \quad (3)$$

Where M is the number of samples, R is the number of principal components retained in the PCA model, and F is the F-distribution with a confidence level 1-α and (R,M-R) degrees of freedom.

Faults detected in the Q-chart are caused by events that break the correlation structure described by the model. An event related to the example in Fig. 1 is the ■ sample where the area of peak 1 is decreased while the area of peak 2 is increased. Under NOC the area of peak 1 and peak 2 are positively correlated, so this event is not described by the model. Hence, there will be a faulty situation only in the Q-chart represented in the residuals of the new sample calculated as:

$$Q_{new} = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2 = \sum_{n=1}^{N} (e_n)^2 \quad (4)$$

where x_n and $\hat{x}_n$ are a new measurement of the nth variable and its predicted (reconstructed) value, respectively, which result in the residual e_n. N denotes the number of variables. There are several ways to determine the confidence limits for the Q-statistic [14,15]. In the present paper, a normal distribution to approximate a weighted χ-square distribution is used from which the UCL for the Q-chart can be calculated according to Jackson & Mudholkar [16]:

$$UCL_Q = \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1-h_0}{\theta_1^2} \right) + \frac{\sqrt{z_\alpha (2\theta_2 h_0^2)}}{\theta_1} \right]^{\frac{1}{h_0}} \quad (5)$$

The matrix $\mathbf{V}$ is defined as the covariance matrix of the residuals $\mathbf{E}$ (after performing PCA on the NOC samples), θ_1 is the trace (the sum of the elements on the main diagonal) of $\mathbf{V}$, θ_2 the trace of $\mathbf{V}^2$, θ_3 the trace of $\mathbf{V}^3$, $h_0 = 1 - ((2\theta_1\theta_3)/(3\theta_1^2))$, and z_α is the standardized normal variable with a $(1-\alpha)$ confidence level.

In standard two-sided SPC charts an observation more than three standard deviations (3σ) from normal operating conditions is often used as the control limit. This corresponds to a coverage probability of 0.9973 ($1 - 2\Phi(-3) = 0.9973$), where $\Phi(\cdot)$ refers to the standard normal distribution operator. In the application described here the

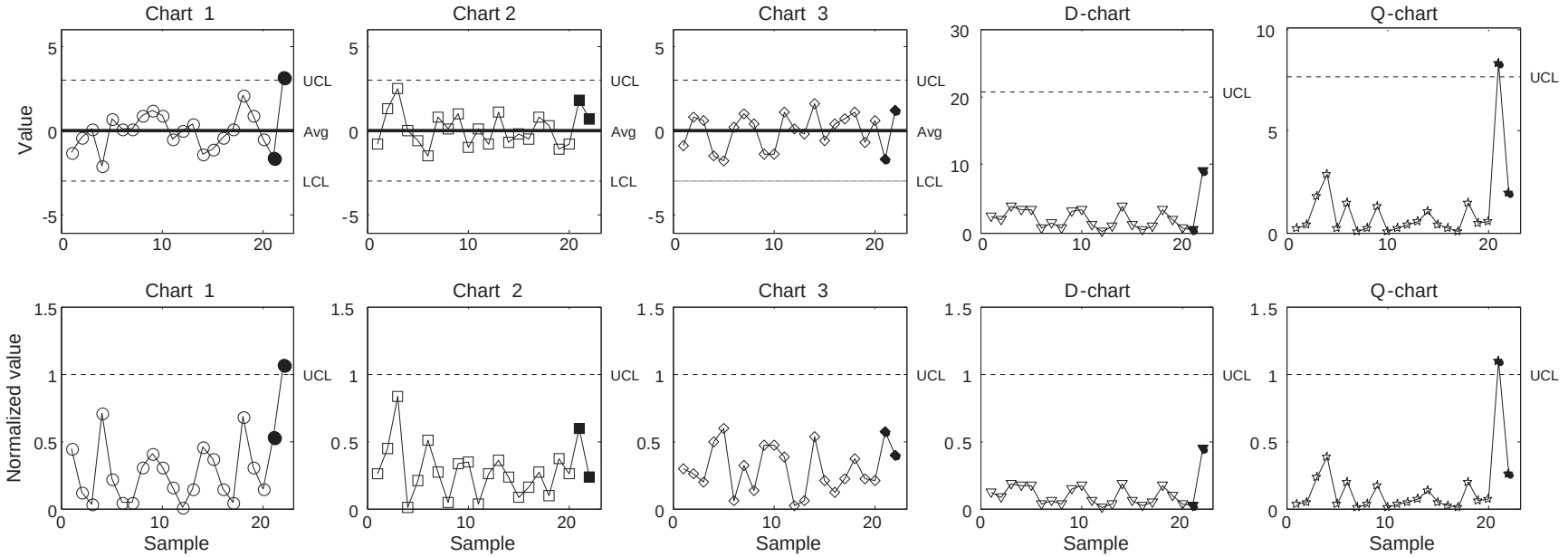

Fig. 2. Simulated example of univariate- and multivariate control charts, before and after normalization with the respective control limit.

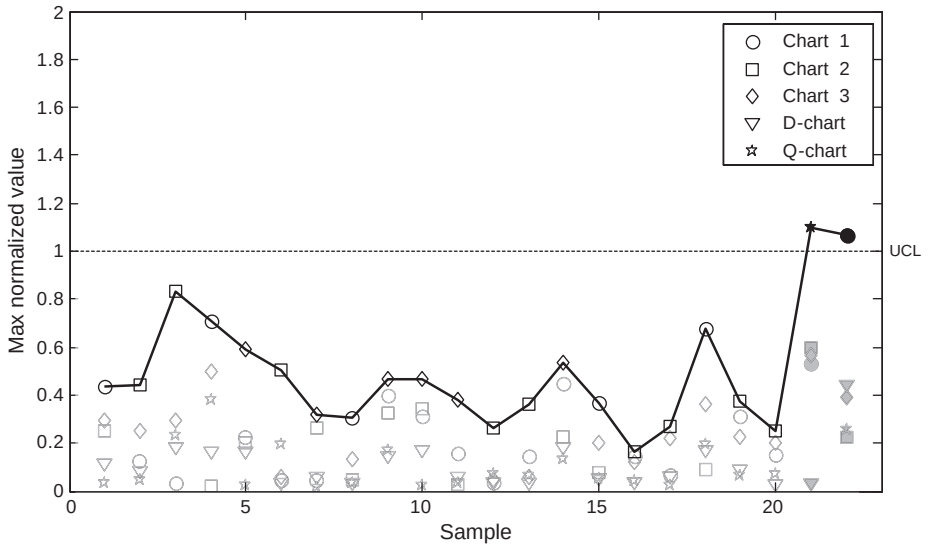

Fig. 3. Simulated example of how univariate- and multivariate control charts can be condensed in one COCO chart monitoring the maximum normalized value for each sample.

one-sided D and Q control chart upper control limit should reflect the same coverage probability, i.e. a 99.73% confidence limit (~3σ) is used.

Note that the use of PCA is under the assumption of a low rank PCA model being adequate. If the variables are independent it suffices to look at the variables individually. The correlation structure in chromatography can be very weak, and if there are only a few (low correlated) variables it would make more sense to use the original approach as proposed by Hotelling [12], instead of using the dimension reduced approach resulting in the D-, and Q-chart. This approach is equivalent to a full rank PCA solution solely evaluating the D-statistics.

For monitoring, the PCA model is applied by projecting a new sample onto the model hyperplane, and calculating the residuals of the PCA model. Then, the associated values of the D- and Q-statistics are calculated for this new sample and the MSPC charts are updated. If a new sample violates the control limit of either statistic the sample is considered to deviate significantly from the samples included in the PCA model, and it is indicative of abnormal process behavior. Once an MSPC chart signals an alarm, the model can be scrutinized to understand the cause of the alarm. One of the most widely used approaches for this is using contribution plots [10,17,18]. Contribution plots compute a list of each single chromatographic peak that contributes numerically to the D- and Q-statistics respectively. However, contribution plots do not automatically reveal the actual reason for the faulty condition. Therefore, those peaks 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 [9].

2.3. Comprehensive control (COCO) charting – monitoring normalized control chart values

The methodology devised here considers both the univariate statistics and the two MSPC statistics (D and Q). Each control chart value is normalized by division with the respective control limit (here 3σ). For the univariate and scaled control values the absolute values and control limits are used. In this way values greater than one indicate deviation from normal operating conditions (NOC), whereas values between zero and one indicate NOC. This is exemplified in Fig. 2 showing three univariate control charts (one for each variable 1, 2, and 3) and the two derived multivariate control charts (D- and Q-chart), before (upper charts) and after normalization (lower charts). The control charts are based on a simulated dataset (autoscaled). The first twenty samples have been used as NOC samples, whereas the two last samples are new independent samples to be monitored. The D- and Q-charts are derived from a two component PCA model explaining approximately 80% of the common-cause variation in the first twenty NOC samples. The control limits corresponds to the 3σ confidence level, and are estimated from the NOC samples.

The two last samples (21 and 22) plotted in Fig. 2 simulate two different types of special causes. Sample 21 is within common-cause variation in all three univariate charts, but clearly deviates in a multivariate sense as it exceeds the control limit in the Q-chart. In contrast sample 22 exceeds the control limit in the univariate Chart 1; in spite of this none of the multivariate charts detects this deviation as being faulty. Consequently, detection of these two different types of

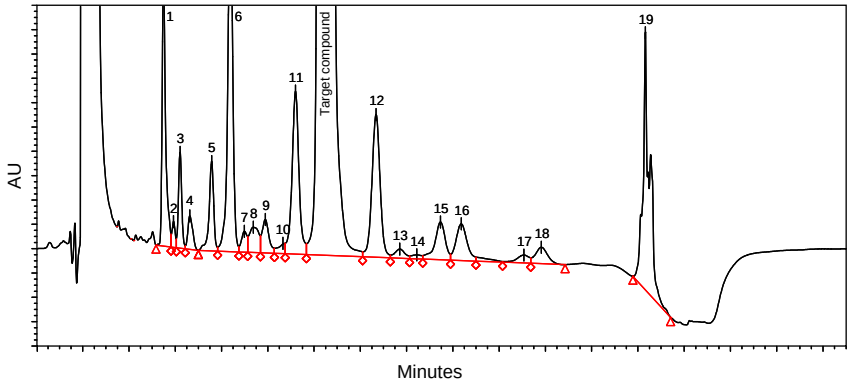

Fig. 4. Analytical chromatogram of a biopharmaceutical drug substance. Selected peaks of interest are marked and integrated.

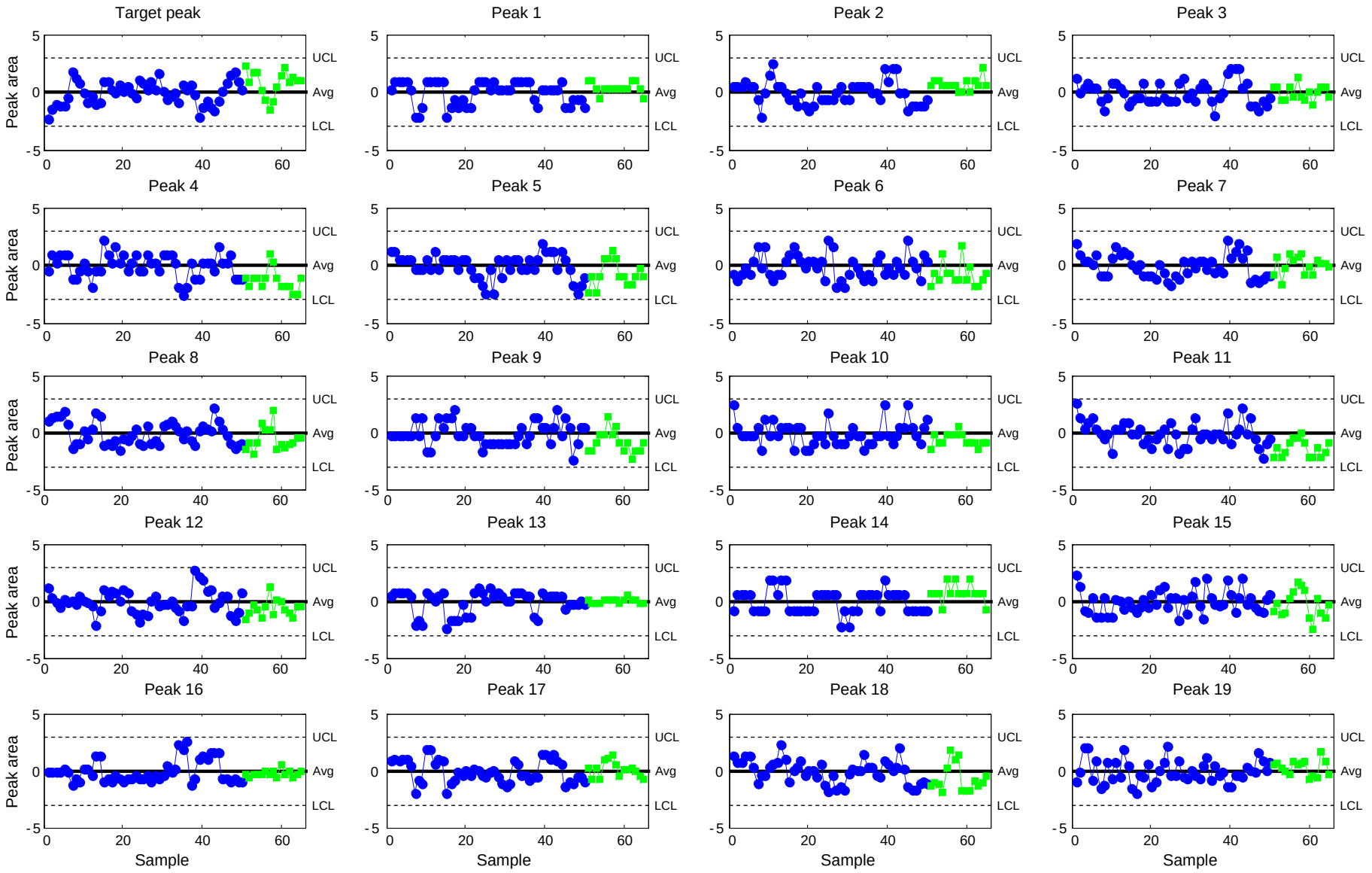

Fig. 5. Univariate control charts for twenty peak areas of calibration set (circle) and validation set (square).

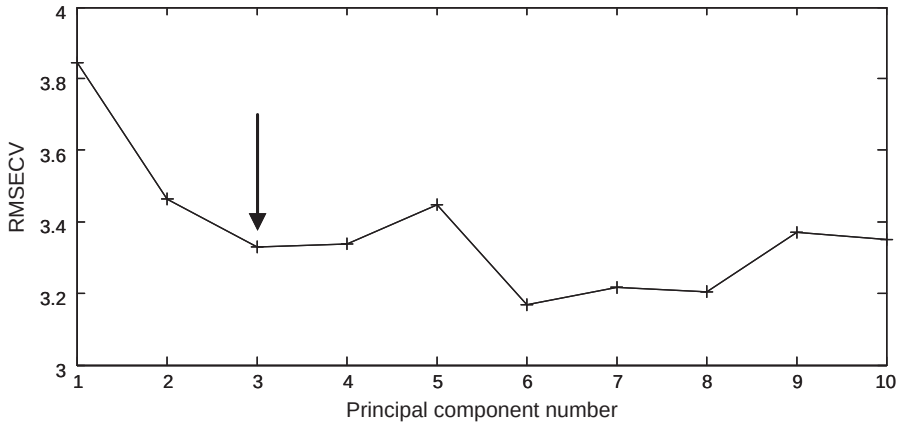

Fig. 6. Results of leave-one-out cross-validation, indicating the optimal number of 3 principal components.

special causes would require inspection of both univariate and multivariate control charts simultaneously. This can be an overwhelming and inefficient task and the risk of missing an out-of-control situation is obvious.

Therefore a more orderly control chart procedure is devised here. As depicted in Fig. 2, Z normalized control values (Z is the number of control charts including the D- and Q-chart, here $Z=5$) is produced for each sample, where the largest value reflects the control chart in which the sample is most deviating. As a condensed measure across all control charts (including the D- and Q-chart) the maximum normalized value is used for COCO charting as exemplified in Fig. 3.

In the COCO chart (Fig. 3) the maximum fault contributions are monitored, allowing both univariate and multivariate statistics to be accounted for at the same time. As opposed to *either* using MSPC *or* using multiple SPC charts, this comprehensive control chart strategy covers the detection capabilities of both.

2.3.1. Estimation of the false positive rate

The presence of false alarms is one of the major reasons that the process operators are skeptical of employing (M)SPC charts for process fault detection. However, it should be noticed that randomly induced false alarms are inherent within (M)SPC. For example, the 99% control limit states that statistically 1% of the normal operating samples will fall outside this limit and incorrectly be identified as faulty (false positive rate). In SPC an observation more than three standard deviations (3σ) from normal operating conditions is often used as the critical limit. 3σ correspond to the upper 0.13% of the distribution $(1-\Phi(3)=0.0013$, where $\Phi(\cdot)$ is the standard normal distribution with mean zero and variance one). As both abnormally high and low deviations are considered, a *single* control chart has a false positive rate of 0.27%. However, if several confidence intervals are considered simultaneously, with coverage probability 0.9973 each, the probability that at least one interval will not contain its true value is greater than 0.0027. Assuming independence between Z control charts the probability of *at least one* of the control charts giving a value greater than one under normal operating conditions can be calculated as: $(1-(1-p)^Z$, with p equal one minus the coverage probability. The independence assumption is indeed *not* valid as the D- and Q-statistics are based on the exact same data generating the univariate statistics. In order not to rely on independence assumptions we device a routine for generic estimation of the false positive rate by bootstrapping the calibration samples [19]. The bootstrap is based on resampling with replacement. Here we generate B different datasets based on a defined calibration sample set (50 samples). These

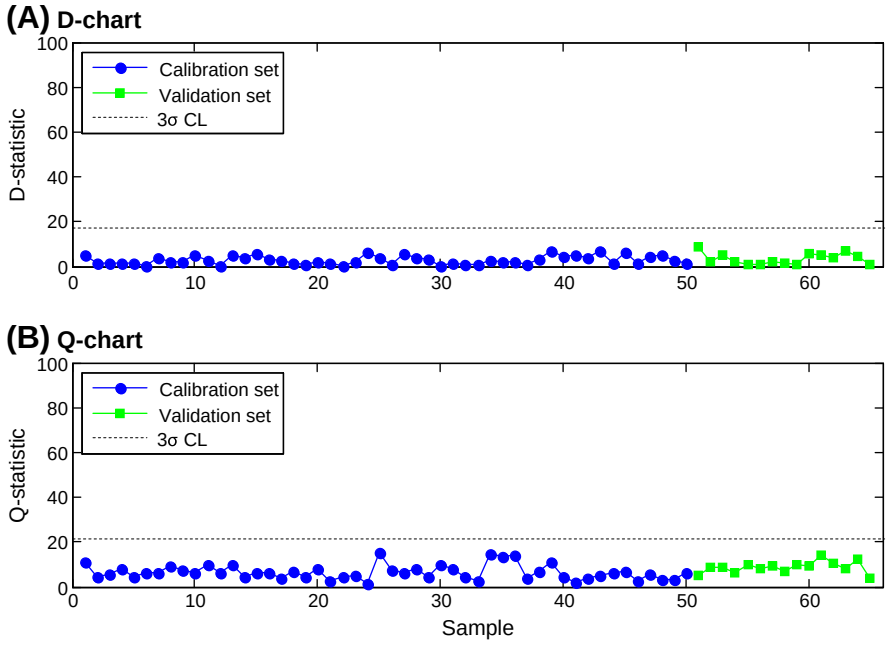

Fig. 7. MSPC charts of (A) D-statistic and (B) Q-statistic of calibration samples (circle) and validation samples (square).

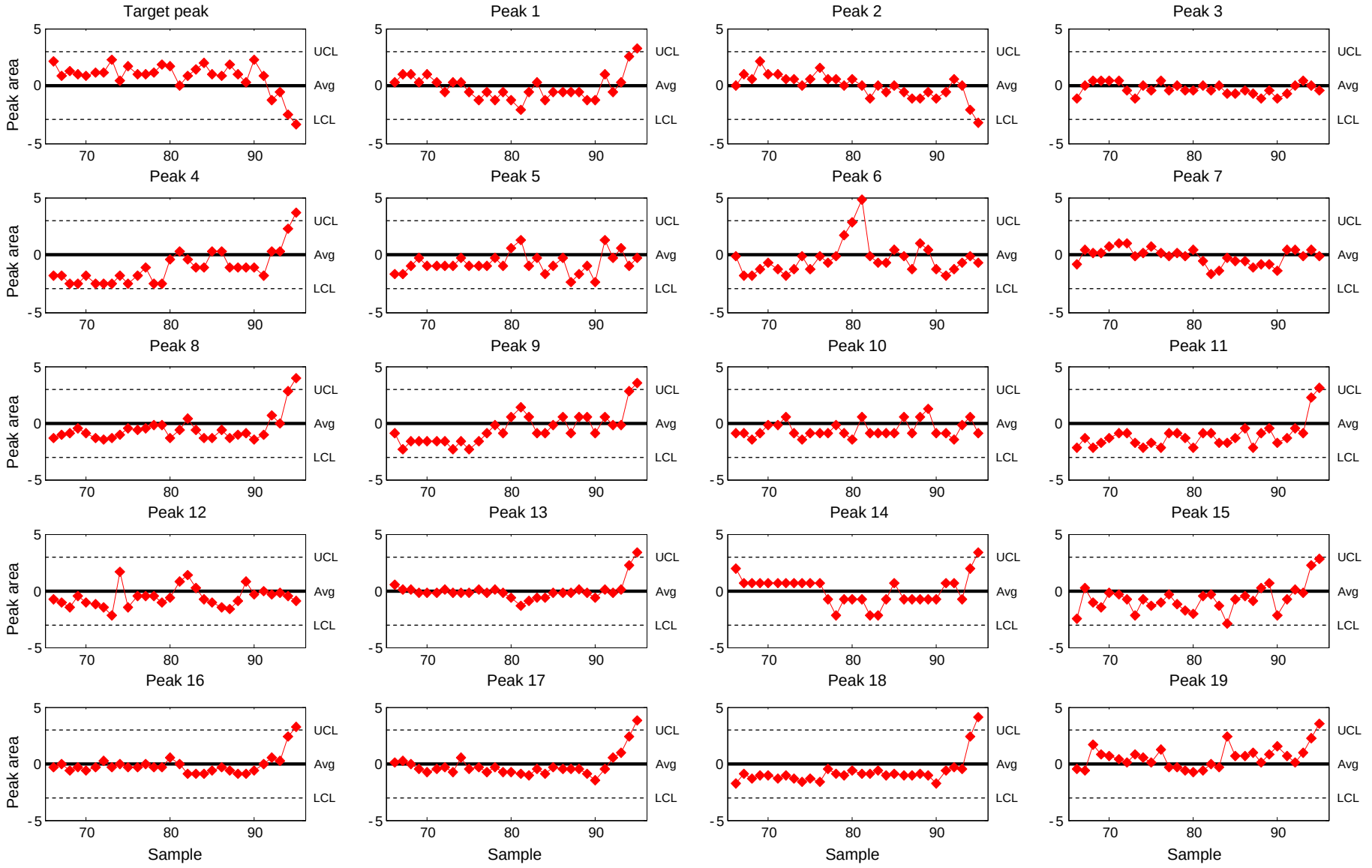

Fig. 8. Univariate control charts for twenty peak areas of thirty test set samples.

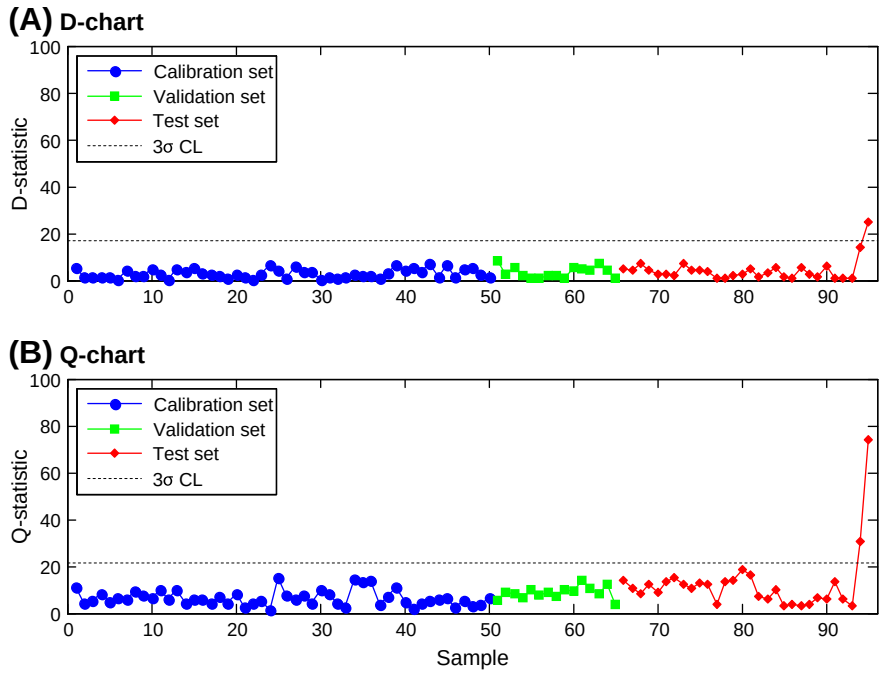

Fig. 9. MSPC charts of (A) D-statistic and (B) Q-statistic of calibration samples (circle), validation samples (square), and test samples (diamond).

are called bootstrap samples and have the same number of samples as the original calibration set. Each of the bootstrap samples is used for building new control charts including estimation of the individual 3σ control limits and estimation of a PCA model and the derived D and Q-statistics with corresponding control limits. A defined validation sample set (15 samples) is then referenced against these control charts. For each validation sample the maximum control chart value is obtained (across all control charts). This is repeated B times. The validation dataset is obtained under NOC and therefore assumed to fall inside the 3σ control limit. The bootstrap procedure is set up to empirically estimate the false positive rate, i.e. the probability of a sample being outside NOC when it is actually a NOC sample. The false positive rate is estimated as the frequency of validation samples obtaining a maximum control value greater than 1 in the COCO chart. The average false positive rate is plotted against B to check for convergence (see Fig. 12). A high false positive rate needs to be accounted for, as otherwise it will result in a loss of confidence in the control chart. A simple way to adjust for the false positive rate is to tune the individual control limits in parallel. The bootstrap approach described above can be applied (with enough iterations) using different control limits producing estimates of the false positive rate as a function of the control limit. Of course this is a trade-off between minimizing the false positive rate without loosing too much sensitivity. However, this is the price when monitoring several parameters simultaneously. For a dataset of the given size (number of variables) it is anticipated that the false positive rate is not going to be detrimental to the ability of the MSPC approach to detect abnormal behavior.

3. Experimental

Ninety-five in-process samples of a high-purity drug substance were collected for routine quality control testing. The first sixty-five samples were collected under NOC, i.e. the process has been running consistently and only high quality products have been obtained. The sixty-five NOC 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 final thirty samples were collected in a process period where forced process changes were applied, giving rise to possible changes in the sample matrix.

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 × 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). The peak areas were integrated and listed in a peak table, and hereafter exported to Matlab version 7 (Matworks, Natick, MA, USA) for further analysis. All software was written in Matlab using tools from PLS_Toolbox (Eigenvector Inc, WA, USA).

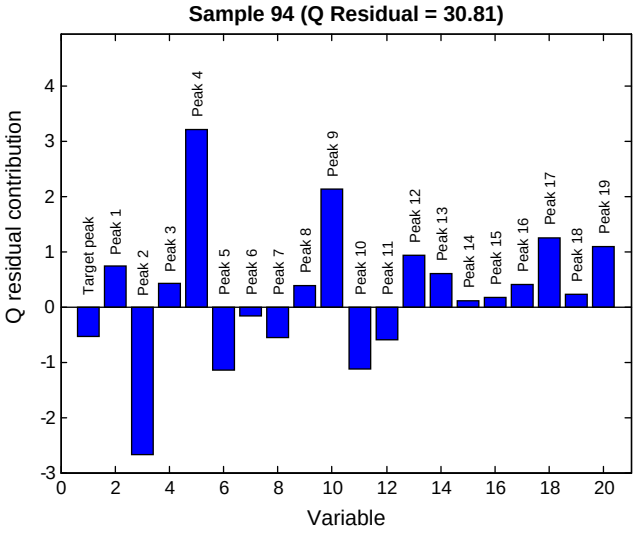

Fig. 10. Q residual contribution plot of the twenty peak areas obtained from sample 94.

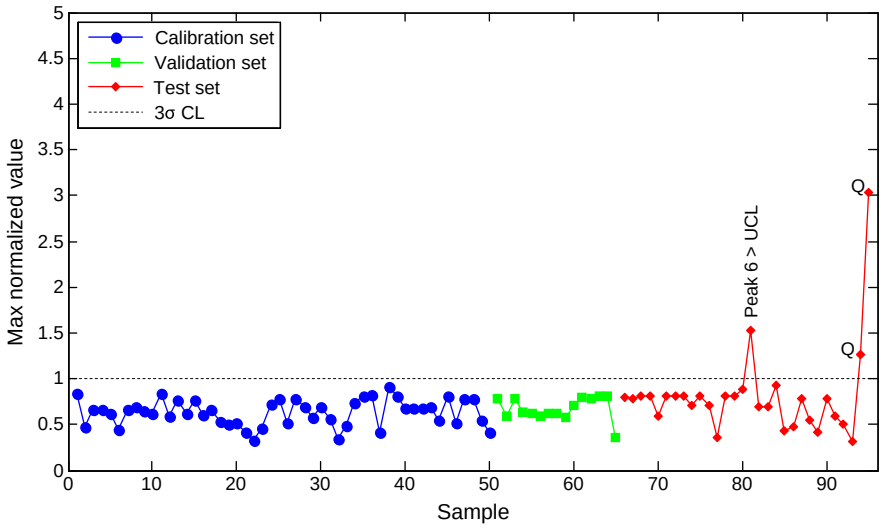

Fig. 11. COCO chart with maximum normalized values of calibration set (circle), validation set (square), and test set (diamond), using a 3σ control limit.

4. Results and discussion

The control chart monitoring can be divided to three distinct phases (initial phase, training phase, application phase). In the first phase (initial phase), historical NOC samples are collected and prepared for modeling. Sixty-five historical HPLC chromatograms obtained for purity analysis of a biopharmaceutical drug substance were collected and the routinely generated peak tables were imported into MATLAB. The peak tables were organized as an $M \times N$ data matrix **X**, with M samples and N peak areas. In Fig. 4 the selected peaks of interest in this study are marked in an analytical chromatogram obtained under NOC.

In addition to the target compound, nineteen impurities are monitored in this study. The 1000-fold difference in concentration for the target compound and most of the impurities is not proportional to the chemical relevance of these compounds. Therefore, all samples were scaled to adjust for the disparity in fold differences, aiming at assuring that all peaks contribute equally to the model.

The essence of the second phase (training phase) is to model the common-cause variation present in the samples obtained under NOC. Since this NOC model exclusively determines whether a new sample is similar or deviates significantly from the NOC samples, the monitoring performance depends very much upon adequacy and representativity of these NOC samples. The number of samples needed to construct an NOC model and control charts depends on the application. In this case study, a calibration set consisting of the first fifty chronologically ordered NOC samples were selected. To validate the model adequacy and representativity of these NOC samples, a validation set consisting of the last fifteen chronologically ordered NOC samples were selected. The autoscaling of the data was based only on the calibration set. Accordingly, the validation set was preprocessed using the parameters determined from the calibration set. In Fig. 5 the scaled peak areas of both the calibration set and the validation set are presented in twenty univariate control charts. The 3σ UCL and LCL are derived from the calibration set data.

By inspection of the twenty univariate control charts presented in Fig. 5, it is observed that all calibration samples are within their respective control limits. Furthermore, the validation samples are all within common-cause variation in a univariate sense. For multivariate monitoring, the calibration set was used to develop a three component PCA model describing 66.14% of the common-cause variation. The selection of an optimal number of three components was based on the results of leave-one-out cross-validation [11] plotted in Fig. 6.

The correlation structure in chromatography can be very weak, thus the number of significant components may be difficult to assess [4]. However, root mean squared error of cross-validation (RMSECV) plotted against PC number in (Fig. 6) has the first clear local minimum at three components, indicating that after this point, the components just reflect noise. The model was validated using the independent validation set consisting of the last 15 chronologically ordered samples. In Fig. 7 the D- and Q-statistics of calibration and validation samples are presented with 3σ control limits derived from the calibration samples.

By inspection of the D- and Q-chart (Fig. 7) it can be confirmed that all sixty-five samples used in the training phase are within the respective 3σ control limits. This confirms that the NOC model represents common cause variation.

In the third phase (application phase) new samples are fitted to the model and monitored using the control charts developed in the training phase. Deviating samples are diagnosed using contribution plots to determine causes of the deviating behavior. The thirty test set samples were collected in a period where forced process disturbances were applied, giving rise to possible changes in the sample matrix. At first the test set samples are monitored in the univariate control charts derived from the calibration samples (Fig. 8).

By inspection of Fig. 8, two different types of special causes are observed. The most notable event is observed for sample 94 and 95, where several impurities increases in parallel, while the target peak

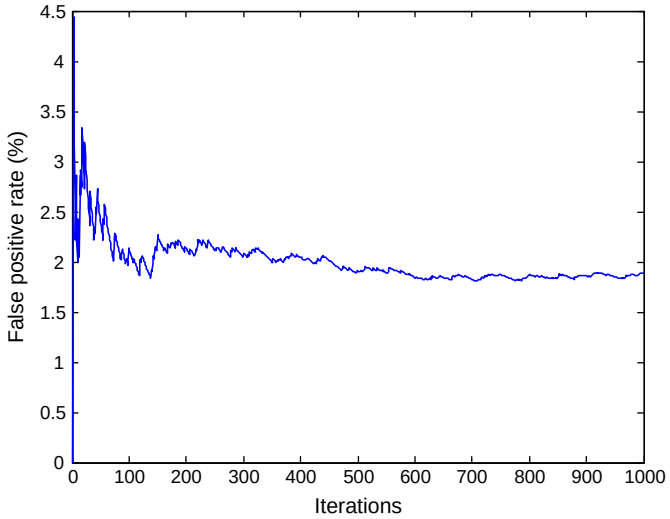

Fig. 12. False positive rate plotted against number of iterations. Convergence is obtained after approximately 600 bootstrap iterations.

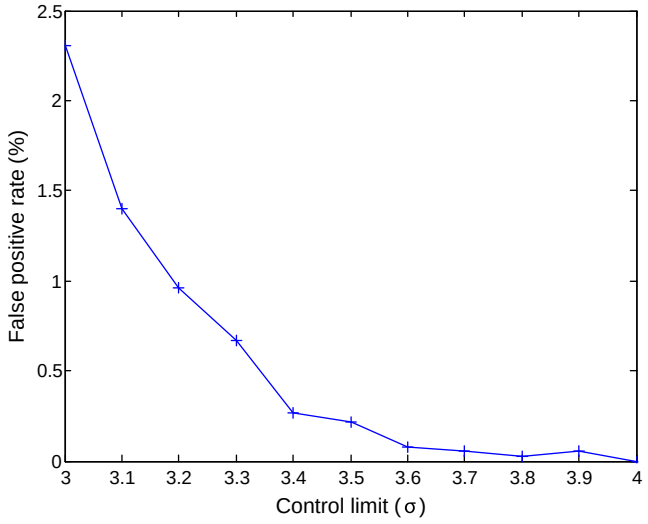

Fig. 13. False positive rate after 600 bootstrap iterations plotted against control limit (σ). Using a 3.5σ control limit will cause the false positive rate to stay below 0.027%.

and peak 2 decreases. In sample 95 several peak areas exceeds their respective control limits, indicating that the process is not running under NOC. The other special cause is observed for impurity peak 6 in sample 81, which as the only peak area exceeds its own control limit.

For multivariate monitoring, the test set samples were exposed to the PCA model, and D- and Q-statistics were derived. As indicated in Fig. 9 only sample 95 is detected in the D-chart, whereas both samples 94 and 95 are detected in the Q-charts exceeding the 3σ control limit.

Sample 94 was not deviating in a univariate sense (Fig. 8) but is deviating in a multivariate sense, and is therefore detected in the Q-chart. To determine chromatographic variables (peak areas) responsible for the signal in the Q-chart, a residual contribution plot is inspected in Fig. 10.

The contribution plot allows us to diagnose the problem with the faulty sample immediately. Clear indications of which peaks that contribute to the deviating behavior are given in Fig. 10. Apparently, this variability is not described by the principal components retained in the NOC model. Accordingly, sample 94 (and sample 95) show up as an abnormal residual variability and a faulty signal in the Q-chart. Sample 81 was previously observed to deviate in a univariate sense as peak 6 exceeded the upper control limit. However, sample 81 is not detected in any of the MSPC charts. This lack of sensitivity of PCA derived statistics is well known, but rarely mentioned. When only a few discrete residuals deviate, the information may potentially be masked when calculating the sum of squared residuals (Q) or T^2 (D). This makes both D and Q non-sensitive measures for monitoring and detection of abnormal situations expressed only in one or a few variables. Therefore, we devise a comprehensive control (COCO) chart that considers both the twenty univariate statistics and the two MSPC statistics (D and Q) as described in Section 2.3. As a condensed measure across all control charts the maximum normalized control value is used for comprehensive monitoring of all samples (Fig. 11).

The devised COCO chart presented in Fig. 11 detects both sample 81 and samples 94–95. Furthermore, the COCO chart indicates the underlying control value causing the faulty signal. In this way comprehensive monitoring of univariate and multivariate information can be conducted, as an overall control chart add-on. However, as described in Section 2.3, inferences likely occur when several control statistics are considered simultaneously, leading to increased false positive rate. Therefore a bootstrap procedure was set up to empirically estimate the false positive rate, i.e. the probability of a sample exceeding the control limit when it is actually a NOC sample. The average false positive rate is plotted against bootstrap iterations to check for convergence in Fig. 12.

The bootstrap exercise presented in Fig. 12 reveals that the false positive rate estimate seems constant just below 2% after approximately 600 iterations. The estimated false positive rate is lower than the theoretical probability of 5.8% when independence between the control values is assumed ($1-(1-p)^Z = 0.0577$, for $p = 0.0027$ and $Z = 22$). This was expected as the D and Q-statistics are based on the twenty univariate statistics. Nevertheless, a false positive rate of approximately 2% may not be acceptable, as too many false warnings will result in a loss of confidence in the control chart and thereby it becomes less effective. Therefore the false positive rate is controlled by tuning the individual control limits in parallel. Of course this will decrease the sensitivity of the COCO chart but preferably makes it more reliable. The bootstrap approach was applied (with 600 iterations) using different control limits producing estimates of the false positive rate as a function of the control limit (Fig. 13).

By inspection of Fig. 13, the estimated false positive rate reaches ~0.2% using 3.5σ as control limit. Selecting the control limit will always be a tradeoff between sensitivity and reliability of the control chart. In this study we aim for a false positive rate below 0.27%, and the 3.5σ control limit was applied to the control chart as illustrated in Fig. 14.

The special cause examples presented in this study are still detected as faulty by the COCO chart after applying a 3.5σ control limit

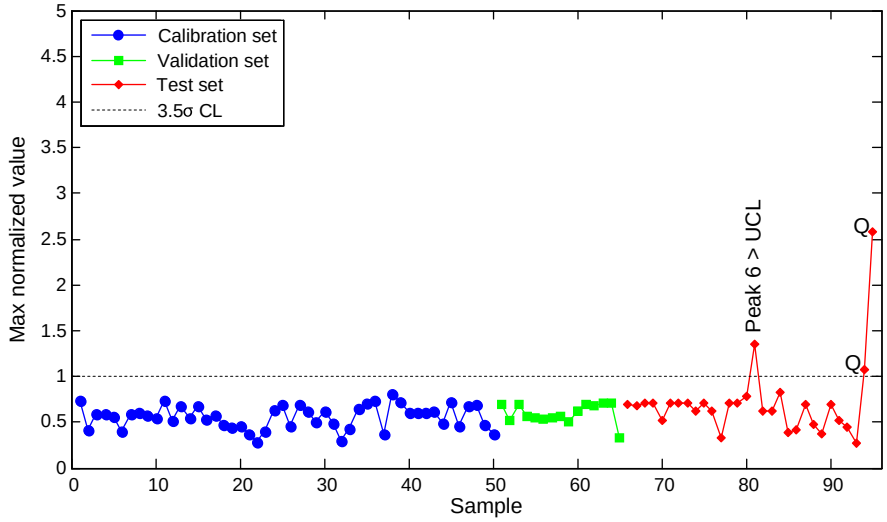

Fig. 14. COCO chart with maximum normalized values, using a 3.5σ control limit.

as indicated in Fig. 14. However, now sample 94 only barely exceeds the control limit as a consequence of the loss of sensitivity.

5. Conclusions and some perspectives

This study demonstrates that MSPC based on PCA can provide early warnings of faulty events in product related analytical chromatography. The study also demonstrates that PCA suffers from lack of sensitivity when faulty events are expressed only in one or a few variables. Therefore a comprehensive control (COCO) chart procedure is devised, that considers both univariate statistics and multivariate statistics derived from PCA in a single condensed plot. This COCO chart allows easy visualization of the combined data from a univariate and multivariate point of view. Two different types of faulty events tested in this study were detected by the COCO chart. However, an increased false positive rate (~2%) was estimated with bootstrapping. This was an expected consequence of inferences occurring when several control statistics are considered simultaneously. The false positive rate was tuned simply by changing the individual control limits in parallel from 3σ to 3.5σ, resulting in a false positive rate below 0.27%. This preferably makes the COCO chart more reliable, at the price of a loss of sensitivity. Applying the COCO chart procedure to multivariate data makes a fully automatic and manageable monitoring possible. Furthermore, if implemented and operated while the chromatographic purity analyses runs, this tool may considerably reduce time needed for subsequent assessment of data, and operate according to the PAT concept aiming for real-time release.

Acknowledgements

The authors thank Casper Leuenhagen and Mads Thaysen (Novo Nordisk A/S, DAPI MDEV) for their helpful suggestions on revising this paper.

References

[1] W. Shewhart, Van Nostrand Reinhold, Princeton, NJ (1931).
[2] D.C. Montgomery, Introduction to Statistical Quality Control, Wiley, NJ, 2005.
[3] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.
[4] A. Nijhuis, S. de Jong, B.G.M. Vandeginste, Chemom. Intell. Lab. Syst. 38 (1997) 51.
[5] A. Nijhuis, S. de Jong, B.G.M. Vandeginste, Chemom. Intell. Lab. Syst. 47 (1999) 107.
[6] 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.
[7] L. Zhu, R.G. Brereton, D.R. Thompson, P.L. Hopkins, R.E.A. Escott, Anal. Chim. Acta 584 (2007) 370.
[8] K. Laursen, S.S. Frederiksen, C. Leuenhagen, R. Bro, J. Chromatogr. A 1217 (2010) 6503.
[9] A. Ferrer, Quality Engineering 19 (2007) 311.
[10] J.F. MacGregor, T. Kourti, Control Eng. Pract. 3 (1995) 403.
[11] S. Wold, Technometrics 20 (1978) 397.
[12] H. Hotelling, in: C. Eisenhart, M.W. Hastey, W.A. Wallis (Eds.), Techniques of Statistical Analysis, McGraw-Hill, New York, 1947, p. 113.
[13] J.E. Jackson, A user's guide to principal components, John Wiley and Sons, 1991.
[14] P. Nomikos, J.F. MacGregor, Technometrics 37 (1995) 41.
[15] S. Joe Qin, J. Chemometrics 17 (2003) 480.
[16] J.E. Jackson, G.S. Mudholkar, Technometrics 21 (1979) 341.
[17] P. Miller, R.E. Swanson, C.E. Heckler, Int. J. Appl. Math. Comp. 8 (1998) 775.
[18] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Chemom. Intell. Lab. Syst. 51 (2000) 95.
[19] R. Wehrens, H. Putter, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 54 (2000) 35.

PAPER II

Laursen K., Frederiksen S.S., Leuenhagen C., Bro R.

Chemometric quality control of chromatographic purity

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

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.

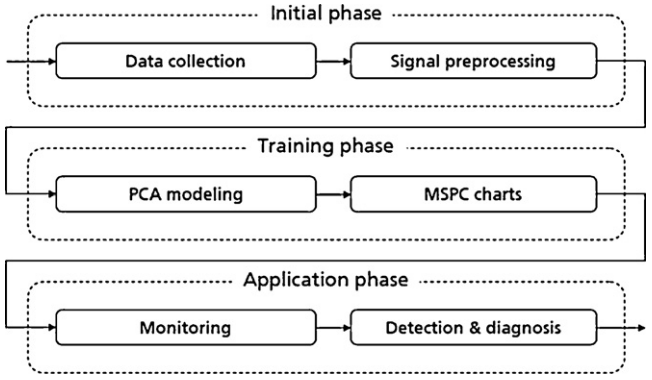

Fig. 1. The three phases according to ITA trajectory (initial, training and application phase).

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-

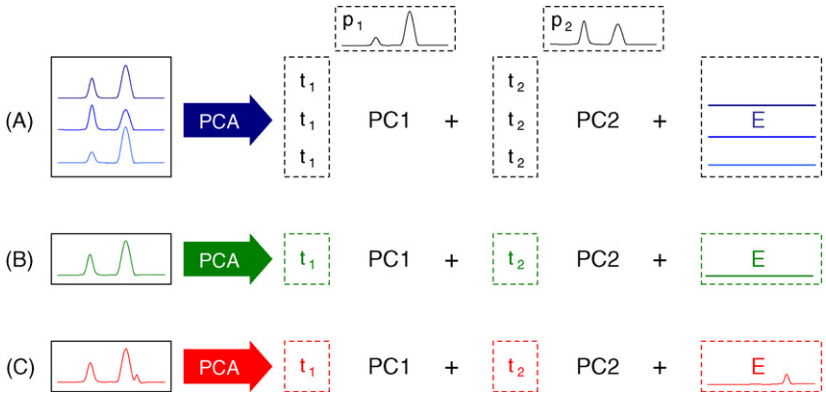

Fig. 2. Example of chromatographic pattern monitoring using PCA. (A) PCA modelling on NOC chromatograms using two principal components. (B) Prediction of a new chromatogram within common-cause variation. (C) Prediction of a new chromatogram deviating from common-cause variation resulting in abnormal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

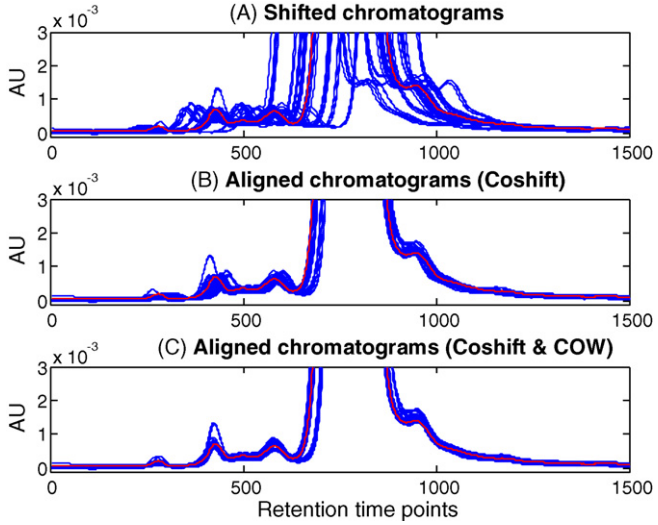

Fig. 3. Plot of shifted (A) and aligned (B and C) chromatograms (blue) towards a reference (red) using Coshift- and COW algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

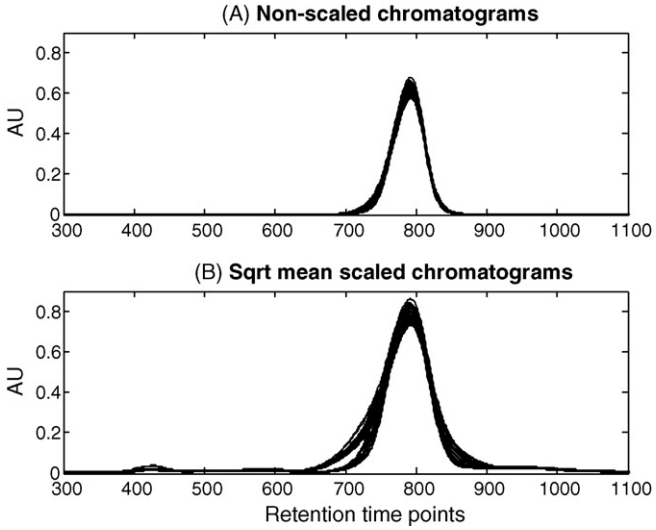

Fig. 4. Plot of chromatograms before (A) and after (B) square root mean scaling.

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

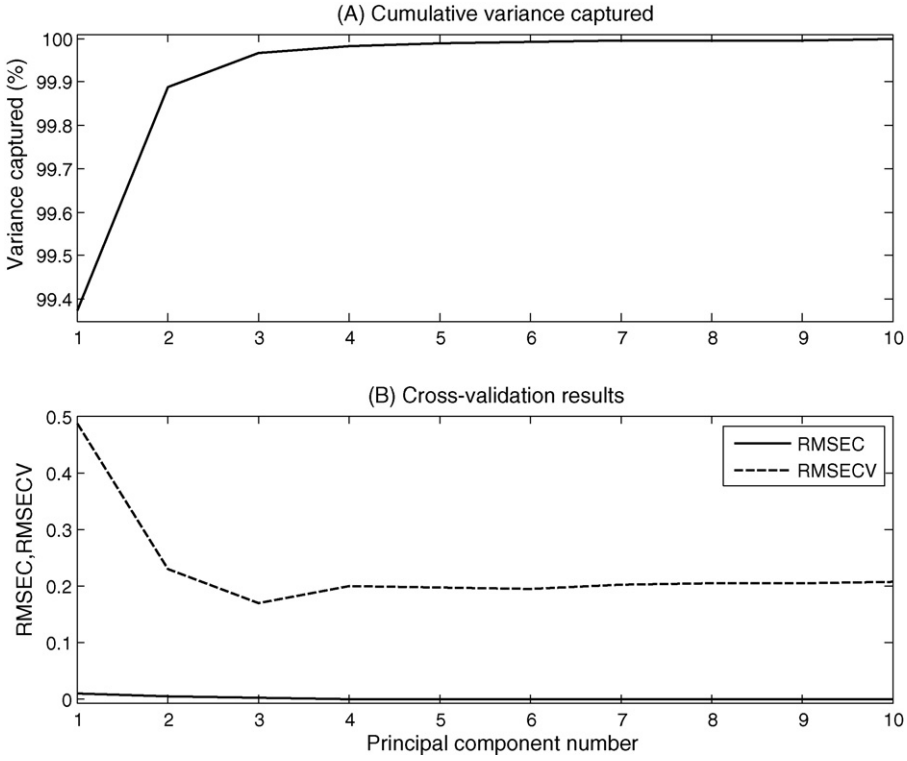

Fig. 5. Plot of cumulative variance captured (A) and results of leave-one-out cross-validation (B).

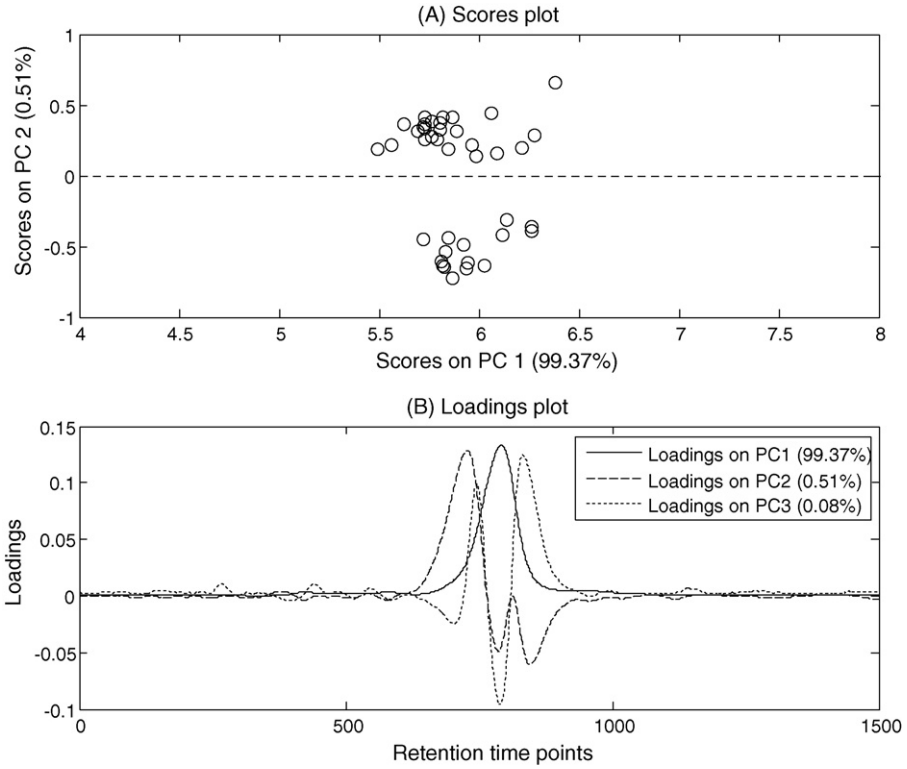

Fig. 6. Scores plot of PC2 vs. PC1 (A) and loadings plot on first three principal components (B).

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

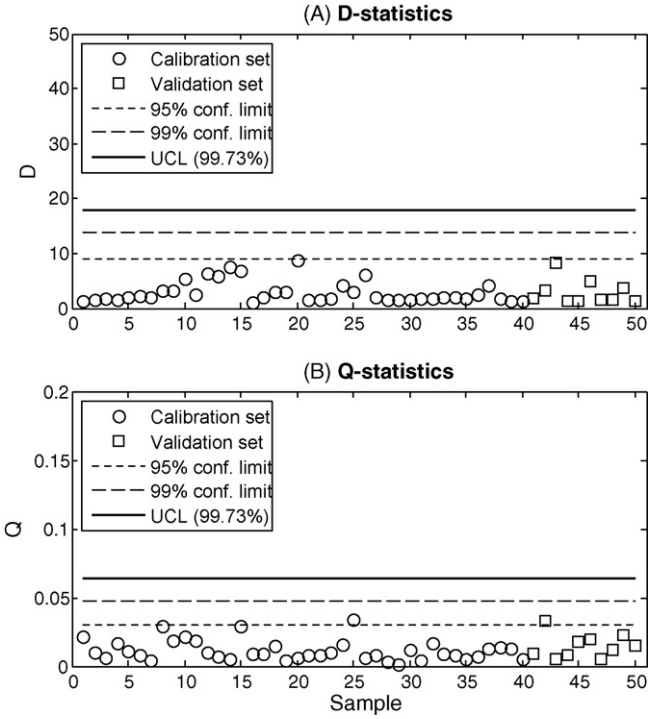

Fig. 7. Plot of (A) D-statistics and (B) Q-statistics of calibration (circle) and validation (square) sample sets.

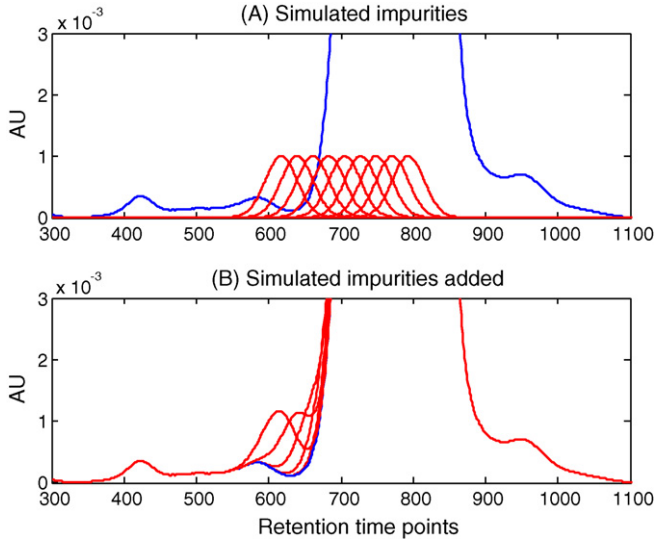

Fig. 8. Simulated 0.1% area impurity peaks (red) in 9 varied resolutions from 0 to 2 before (A) and after (B) added to a reference chromatogram (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

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

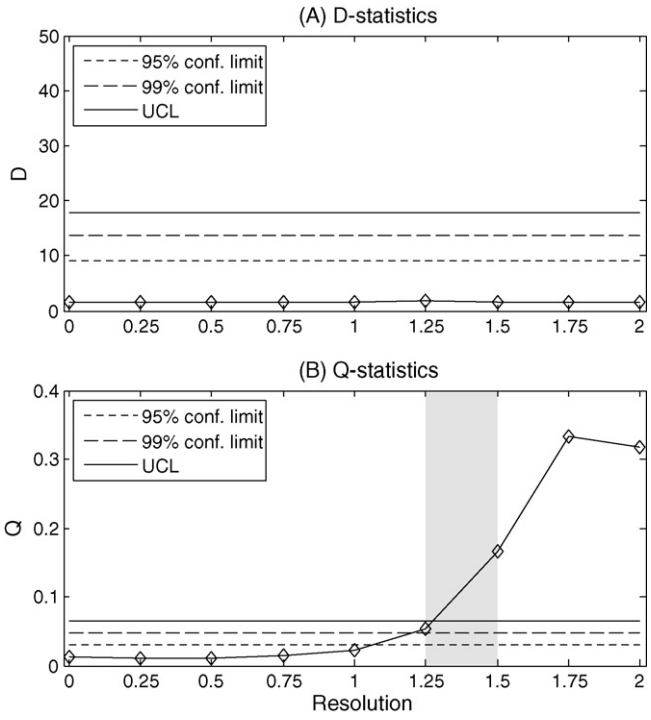

Fig. 9. Plot of D-statistics (A) and Q-statistics (B) of chromatograms added 0.1% area impurity with varying resolution (R_s 0–2). Critical area of detection in Q-statistics is marked.

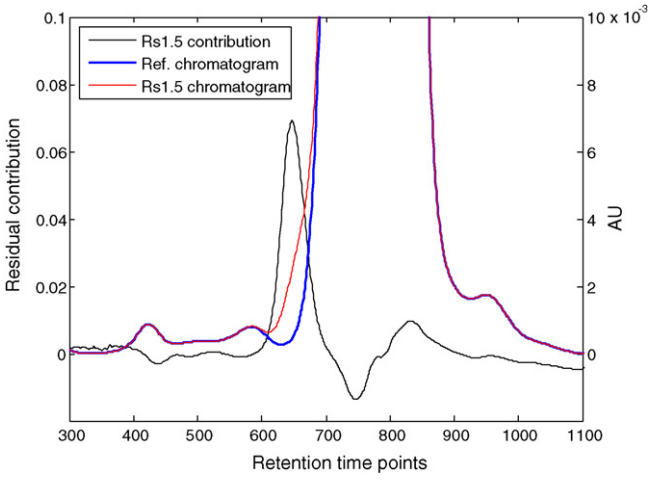

Fig. 10. Plot of the faulty R_s 1.5 residual contribution (black), plotted together with the reference (blue) and the faulty R_s 1.5 chromatogram (red) on the secondary *y*-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

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.

PAPER III

Laursen K., Justesen U., Rasmussen M.A.

Enhanced monitoring of biopharmaceutical product purity using liquid chromatography - mass spectrometry

Journal of Chromatography A, 1218 (2011) 4340-4348

Journal of Chromatography A, 1218 (2011) 4340–4348

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Enhanced monitoring of biopharmaceutical product purity using liquid chromatography–mass spectrometry

Kristoffer Laursen[a,b,*], Ulla Justesen[b], Morten A. Rasmussen[a]

[a] 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 8 December 2010
Received in revised form 15 February 2011
Accepted 26 April 2011
Available online 6 May 2011

Keywords:
LC–MS
Impurity detection
Principal component analysis (PCA)
Multivariate statistical process control (MSPC)
Multiple testing
Signal preprocessing

ABSTRACT

LC–MS is a widely used technique for impurity detection and identification. It is very informative and generates huge amounts of data. However, the relevant chemical information may not be directly accessible from the raw data map, particularly in reference to applications where unknown impurities are to be detected. This study demonstrates that multivariate statistical process control (MSPC) based on principal component analysis (PCA) in conjunction with multiple testing is very powerful for comprehensive monitoring and detection of an unknown and co-eluting impurity measured with liquid chromatography–mass spectrometry (LC–MS). It is demonstrated how a spiked impurity present at low concentrations (0.05% (w/w)) is detected and further how the contribution plot provides clear diagnostics of the unknown impurity. This tool makes a fully automatic monitoring of LC–MS data possible, where only relevant areas in the LC–MS data are highlighted for further interpretation.

1. Introduction

Analytical monitoring of impurity profiles in biopharmaceutical products (drug substances and drug products) is important for tracking the product quality. Impurities may potentially have adverse effects and must be identified, qualified, and reported according to the respective thresholds [1,2]. Increasing demands for higher biopharmaceutical product quality has been facilitated by developments in analytical instrumentation and computer systems. This trend leads to new and better tools for monitoring, detection, and identification of new impurities in a timely fashion.

Analytical separation techniques based on high performance liquid chromatography (HPLC) with UV detection are commonly used for determination of impurities in biopharmaceutical products. The separation and subsequent detection of compounds in a sample delivers a chromatogram, which ideally allows separation of peaks which can be attributed to individual chemical compounds. For high-purity drugs, the target compound is present in excess compared to a potential impurity. Hence, detecting the occurrence of an unknown impurity co-eluting with the target compound is a particular problematic challenge. Therefore, purity analysis of a biopharmaceutical product often entails purity examination of the target peak. Peak-purity examination should prevent co-eluting impurities to escape detection in the conventional HPLC analysis [3].

HPLC with diode array detection (HPLC-DAD) is a commonly used method to conduct peak-purity examination. However, many impurities are structurally related to the drug substance, and their structure contains very similar chromophores, making purity assessment based solely on HPLC-DAD data difficult and unreliable. Coupling a mass spectrometer to a liquid chromatograph (LC–MS) brings more selective signals to the table. LC–MS is probably the most powerful technique currently available for pharmaceutical analysis [4]. The technique is still under fast development, particularly in the mass spectrometry area, with vastly improved sensitivity and resolution. However, such state-of-the-art high-resolution instruments are considered rather costly for routine analysis in a pharmaceutical manufacturing environment. Moreover, these high-resolution LC–MS instruments may not contribute with additional required information compared to conventional low cost LC–MS instruments. Since a mass spectrometer (MS) separates compounds by their respective mass-to-charge ratios (m/z), any difference in the m/z values between the impurities and the drug substance will allow an unambiguous detection regardless of similarities in their UV spectra. Therefore an impurity co-eluting with the target peak will be separated in MS as long as their m/z values are different and ionization of the impurity is not suppressed by the target compound. The LC–MS technique is very informative and generates huge amounts of so-called three-way data, where

* 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 © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.04.080

each sample is characterized by the intensity as a function of retention time and m/z. However, the relevant information from the chemical point of view is not directly accessible from the raw data map, which makes manual interpretation tedious and often generates a bottleneck in the analysis process [5]. Furthermore manual inspection of LC–MS data is prone to subjective decision-making likely to cause additional errors. Several advanced techniques for the assessment of LC–MS peak purity and co-elution problems have been reported during the last decades [6–9]. However, to comply with increased focus on process analytical technology (PAT) and quality by design (QbD) there is a need for an automatic tool that routinely monitors, detects, and extracts relevant signals from the LC–MS data where further interpretation and identification should be focused. Furthermore, such a tool should detect relevant variation in the LC–MS map quantitatively and in a statistically reliable way. This is a relatively unexplored area in LC–MS data analysis. A powerful tool has recently been demonstrated on chromatographic purity analysis by Laursen et al. [10]. That study demonstrates that multivariate statistical process control (MSPC) based on principal component analysis (PCA) [11,12] applied on chromatographic data is suitable for monitoring subtle changes in the chromatographic pattern. Unknown impurities co-eluting with the target compound were detected in the sum of squared residuals (Q) statistics, and contribution plots provided clear diagnostics of cause of the subtly deviating chromatograms [10]. However, this approach might suffer from lack of sensitivity when applied to LC–MS data. The huge amount of data points combined with the discrete nature of LC–MS signals (i.e. sharp signals in MS direction) makes detection of unknown impurities a case of needle-in-the-haystack expedition. If a new LC–MS sample containing an unknown impurity is fitted to a PCA model based on normal operation condition (NOC) LC–MS samples, the resulting residuals would ideally hold information about the unknown impurity. However, a few discrete residuals related to an unknown impurity would simply be masked when calculating the sum of squared residuals (Q). This makes Q a non-sensitive measure for monitoring and detection of unknown impurities present in low concentrations. Therefore, a more discriminative and sensitive measure is needed targeted towards the nature of LC–MS data. Ralston et al. [13] proposed a statistical enhancement to the typical application of multivariate statistical techniques. The statistical enhancement uses confidence limits on the residuals of each variable for fault detection rather than just confidence limits on the overall Q residual. The method detected faults earlier than the basic Q residual contribution method typically used, but the enhancement proved primarily as a graphical support tool and not as a single value measure for control chart monitoring.

In this study, the approach reported by Laursen et al. [10] is developed to adapt the nature of LC–MS data and to enhance monitoring and detection of unknown impurities in an industrial insulin intermediate (DesB30). In-process samples are spiked with the structurally related human insulin drug product co-eluting with DesB30. MSPC based on PCA is combined with variable wise (multiple) testing. This would enhance detection of discrete residuals from unknown impurities, as residuals of each variable are tested against corresponding model residuals.

2. Theory and methods

The general workflow of MSPC based on PCA follows a previously described trajectory [10,14]. The trajectory is divided in three phases; the initial phase, the training phase and the application phase (ITA). In this modified version, the training phase involving PCA modeling is extended with multiple testing as shown in Fig. 1.

In the initial phase, appropriate historical LC–MS experiments are collected and prepared for PCA modeling. In the training phase a PCA model based on NOC LC–MS samples is developed (describing common cause variation) and multiple testing is applied on the residuals. Finally, in the application phase new samples are fitted to the model and the most significant variable is monitored in control charts developed in the training phase. Deviating samples are diagnosed using multiple testing contribution plots to determine causes of the deviating behavior.

2.1. Signal preprocessing

Once the LC–MS data has been collected, preprocessing methods are required to correct, refine and filter the data. The quality of signal preprocessing is crucial in order to extract relevant (chemical) information. The signal preprocessing was divided into the following steps: baseline correction, normalization, alignment, data reduction, and scaling. The preprocessing steps are described in the following subsections. The practical implications of these preprocessing steps are visualized in the result section.

2.1.1. Baseline correction

Baseline correction is commonly employed to eliminate interferences due to baseline drift. A variety of techniques for baseline correction of LC–MS data are applicable and is reviewed by Listgarten and Emili [15] among others. In this study an efficient and rather simple method for baseline correction is applied. The method works by fitting a global polynomial (of a user-defined order) to each extracted ion chromatogram and, through an iterative routine, down-weighting points belonging to the signal. A baseline is then constructed and subtracted from the original extracted ion chromatogram. Upon selecting the polynomial order and fraction of data points to use for determining the baseline, the algorithm provides an objective and automatic preprocessing. The baseline correction is similar to a previously described method by Gan et al. [16].

2.1.2. Normalization

MS signals are frequently corrupted by either systematic or sporadic changes in abundance measurements. Normalization will correct for bias due to errors in sample amount, possibly sample carry-over and drifts in ionization and detector efficiencies. Normalization procedures enable a more accurate matching and quantification between multiple samples. Different procedures for normalization can be applied. Normalization values can be calculated on the basis of a global distribution for all detected features (like sum, average or median of all intensities per run), or calculated from a specific sub-set of features, for instance from a spiked protein that is used as internal standard [15,17]. In this application the target peak purity might vary but the overall signal intensity should ideally be the same for each sample. Therefore the sum of all intensities is used as normalization value for each sample.

2.1.3. Alignment

As with every laboratory experiment, chromatographic separation is stable and reproducible only to a certain extent. The retention time often shows large shifts, and distortions can be observed when different runs are compared. Even the m/z dimension might show (typically much smaller) deviations. Pressure fluctuations or changes in column temperature or mobile phase may result in shifted peaks.

Alignment of shifted peaks can be performed in various ways. Very reproducible LC–MS data often need only a movement of the extracted ion chromatograms a certain integer sideways for proper alignment. This is characterized by a systematic shift and can easily be handled by the recently published *i*coshift algorithm [18].

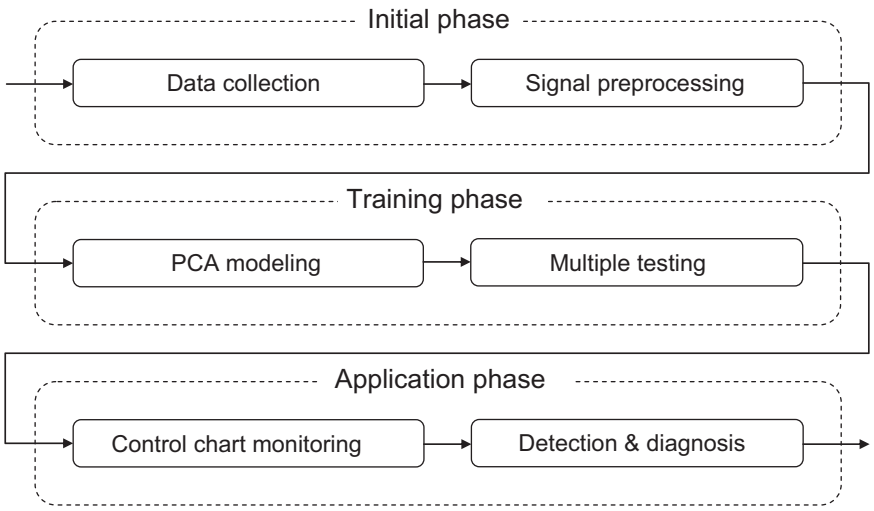

Fig. 1. The three phases according to ITA trajectory (initial, training and application phase).

The *i*coshift algorithm is based on correlation shifting of intervals and employs a fast Fourier transform engine that aligns all spectra simultaneously. The algorithm is demonstrated to be faster than similar methods found in the literature making full-resolution alignment of large datasets feasible [18]. Yet, if peaks shift independently from one another in the same extracted ion chromatogram, more complex shift correction is needed to correct for this non-systematic shift [19,20].

2.1.4. Data reduction

The LC–MS map of a sample is characterized by a collection of intensity measurements as a function of retention time and m/z value. To make the measurements more comparable, and to reduce the huge amount of data points per sample, all intensities within a user-specified bin level are summed. This technique puts all the intensities on a (time, m/z) grid. The bin size is selected based on experience.

2.1.5. Scaling and centering

Scaling is crucial for the performance of the subsequent multivariate statistical analysis. A fold difference in concentration for the target compound and an impurity is not proportional to the chemical relevance of these compounds [21]. Therefore scaling is applied to increase the model sensitivity on detecting small unknown impurities. Furthermore, scaling is crucial in order to bring the distribution of data points close to a normal distribution. This is especially important when multiple testing (like Student's t-test) is used for difference analysis [22]. In many cases, a logarithmic transformation is used for stabilization of the variance. Furthermore, using log-transformed intensities, the disparity in fold differences in between various signals is adjusted. As the final preprocessing step the samples are mean centered (the average unfolded chromatographic pattern is subtracted) to remove a common offset. This brings each variable to vary around zero. This procedure is standard in multivariate modeling that focuses on variability in data.

2.2. MSPC based on PCA modeling combined with multiple testing

PCA and variable wise (multiple) testing offers two different dimensions to statistical data analysis. Multiple testing aims at separating the variable space into variables with a significant- or non-significant change, where PCA separates data into a systematic part (D) and a non-systematic part (Q). In Fig. 2 this is schematized.

Experiments where a high number of variables are evaluated on possibly several outcomes involve testing of numerous hypotheses where handling of error rates is of crucial importance. This discipline is referred to as multiple testing. Multiple testing is widely used for biomarker discovery in proteomics, and has been applied in several difference analyses of LC–MS data intensities [15,23,24]. Both Wiener et al. [23] and Listgarten et al. [24] evaluate the intensity differences between samples from two classes using t-tests on every combination of time and mass to charge ratio, to find regions of interest for further interpretation. However if multiple testing is applied directly to preprocessed LC–MS data it would result in detection of all intensity differences (i.e. both known according to normal operating conditions and unknown features). Multiple testing applied to PCA residuals would only result in detection of unknown features, as the known features are described by the model and expressed in the D-statistics.

With PCA the variation from many correlated (time, m/z) bins in a data matrix $\mathbf{X}$ (with M rows of samples and N columns of bins), can be decomposed into R ($R \leq N$) linear principal components $\mathbf{TP}^{\mathrm{T}}$ and a residual part $\mathbf{E}$ ($M \times N$):

$$\mathbf{X} = t_1 {p_1}^T + t_2 {p_2}^T + \cdots + t_R {p_R}^T + \mathbf{E} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \quad (1)$$

where $\mathbf{T}$ ($M \times R$) is the score matrix and $\mathbf{P}$ ($N \times R$) is the loading matrix, with R components. $\hat{\mathbf{X}}$ is the matrix of predicted values. The correct number of significant principal components can be determined by using cross-validation to eliminate less important directions in the data matrix [25]. In this way the dimensionality of the data matrix is reduced while capturing the underlying relationship between the variables. In standard PCA, each sample is a vector of values. If one sample is a matrix of values (e.g. in the case of LC–MS data), the sample matrix can be unfolded into a vector. This allows standard application of PCA, but throws away some of the information conveyed by storage in a matrix. Using the information contained in all the measured signals simultaneously, MSPC charts are much more powerful in detecting faulty conditions than conventional single variable SPC charts [26]. Once the MSPC chart signals an alarm, the model can be scrutinized to understand the cause of the alarm; hereafter a possible corrective action can be taken. Faults can be due to deviation from *common-cause variation* (detected in Q) and in the *magnitude* of the common cause variation (detected in D). Fault detected in the D chart could for example be caused by an increased amount of already modeled compounds in the sample, and is described by the scores in

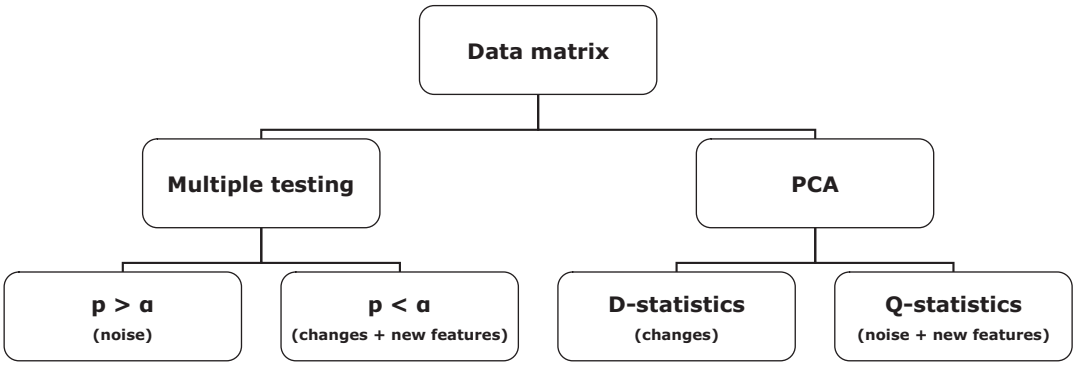

Fig. 2. Schematic overview of two different data analytical approaches for extraction of information from multivariate data. p refers to test probability, α is significance level.

Hotelling's T^2. Hotelling [27] introduced the T^2 for principal components, also referred to as D:

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

where t_r is the rth principal component score, $\sigma_{t_r}^2$ is the variance of the rth component and R denotes the number of principal components retained in the PCA model. Assuming normality for the individual scores, the D-statistic can be expected to approximately follow a weighted F distribution and the upper control limit for the D-statistic can be calculated according to Jackson [28].

If a new sample, containing an unknown impurity, is predicted by the model (based on pure samples), the sample is expected to break the correlation. Indications of the unknown impurity would then be represented in the residuals and monitored in Q:

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

where x_n and $\hat{x}_n$ are a measurement of the nth variable and its predicted (reconstructed) value, respectively which result in the residual e_n. N denotes the number of variables. Most commonly, a normal distribution to approximate a weighted Chi-square distribution is used from which the upper control limit for the Q-statistic can be calculated according to Jackson and Mudholkar [29].

However, as claimed earlier, a few discrete residuals related to an unknown impurity would simply drown when calculating Q. In order to detect the needle in the haystack we device multiple testing based on a simple t-test for each bin (n) as:

$$t_n = \frac{e_{new,n} - \bar{e}_{ref,n}}{s_n \cdot \sqrt{1 + M^{-1}}} \tag{4}$$

where

$$s_n^2 = \frac{1}{M-1} \sum_{i=1}^{M} (e_{i,n} - \bar{e}_{ref,n})^2 \tag{5}$$

and

$$\bar{e}_{ref,n} = \frac{1}{M} \sum_{i=1}^{M} e_{i,n} \tag{6}$$

where $e_{new,n}$ is the residual from the new sample for bin n, $\bar{e}_{ref,n}$ is the mean of the residuals from the reference samples for bin n. M is the number of reference samples. s_n is the standard deviation of residuals from reference samples for bin n.

The critical value of t is dependent on sample size. In order to correct for this ambiguity t is transformed to a z-value through a p-value:

$$P(T_{df} \leq t_n) = \Phi(z_n) \tag{7}$$

where T_{df} is the t-distribution with df degrees of freedom, $df = M - 1$. Φ is the cumulative distribution function of the standard Gaussian distribution. This z-value is used as diagnostic measure for the corresponding (time, m/z) bin. The z-value and p-value reflects the same statistics (Eq. (7)) and hence the behavior of the system. When dealing with signals of interest in the area of $p < 0.01$, changes are more easily captured by exploring the corresponding z-values e.g. over production time.

2.2.1. Multiple testing

Handling of issues related to multiple testing is becoming more important as number of features detectable from modern analytical instruments is rapidly increasing. For example within the field of proteomics from different platforms such as micro arrays, LC–MS, GC–MS, and NMR often numbers in thousands to tens of thousands or even more is common [30]. Performing numerous univariate significance tests on such highly multivariate data will lead to a high false positive rate (FPR). The conservative Bonferroni factor is a way of controlling the error rate across all tests, known as the family wise error rate (FWER) [31]. The Bonferroni factor is simply a proportionality correction of the p-value threshold (α) with the inverse of the number of test. The Bonferroni correction is a crude up front correction where *all* null hypotheses are assumed true i.e. no difference what so ever. But data is seldom collected under the assumption that there is no relation with a specified outcome. In 1995 Benjamini and Hochberg [31] developed control of false discovery rate (FDR) as an alternative to Bonferroni factor in multiple testing. Estimation of the FDR, contrary to FWER, does not assume that all null hypotheses are true but estimates the proportion of null cases and non-null cases from data. This procedure is shown more powerful in detecting true non-null cases than procedures controlling the FWER [32]. Where the FPR predicts how many of the truly null hypotheses are rejected, the FDR predicts how many of the rejected hypotheses are in fact likely to be truly null. In proteomics the aim is to discover biomarkers in order to develop biological understanding. Here a list of significant biomarkers supported by a FDR is relevant for reporting of results including statistical inference. In MSPC the primary scope is to deem a sample pure or impure and secondly if impure to investigate the impurity contribution. Both cases are dealing with issues related to multiple testing, but as the scope is different, the estimation and extraction of a relevant statistics is likewise. In the following subsection we derive a single measure statistics, and estimate its distribution under normal operator conditions.

2.2.2. Single measure statistic for control chart

In Laursen et al. [10] the *Q* value was used for a new sample as a measure for detecting subtle differences in the chromatographic pattern. The methodology devised here produces not one but *N* significance tests where *N* is the number of bins. These are expressed as a list of *z*-values; $z_1, z_2, \ldots, z_K$. The largest values of $z_1, z_2, \ldots, z_K$ reflect the bins where the new sample is most deviating. Impurities are in excess and hence only large positive *z*-values are of interest. The present method proposes use of the *maximum* *z*-value across all *K* bins as a measure in control chart monitoring.

2.2.3. Distribution of the maximum z-value across N bins

Under normality assumptions for residuals within each bin, with equal variance for calibration and new samples, the derived *t*-test statistics is *T* distributed with $M-1$ degrees of freedom (*M* number of calibration samples). The corresponding *z*-values are normally distributed with mean zero and variance one. Assuming independence between the *K z*-values it is easy to compute the distribution of the maximum *z*-value:

$$P(z_{\max} \le z) = P(z_1 \le z) \cdot P(z_2 \le z), \ldots$$

$$, P(z_K \le z) = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt \right)^K \quad (8)$$

In standard two-sided SPC charts an observation more than three standard deviations (3σ) from normal operating conditions is often used as the critical limit. This correspond to a coverage probability of 0.9973 ($1 - 2\Phi(-3) = 0.9973$). As only maximum positive *z*-values are of interest here, the one-sided control chart threshold should reflect the same coverage probability. In accordance it is possible to calculate the corresponding threshold for the maximum *z*-values ($z_{0.9973}$) such that $P(z_{max,K} \le z_{0.9973}) = 2\Phi(-3)$. This threshold only depends on number of bins (*N*). For $N = 1000$, $z_{0.9973} = 4.55$, and for $N = 500$, $z_{0.9973} = 4.40$. Independence between bins might be an overly optimistic assumption, especially when chemical compounds give signal in more than one bin. In order not to rely on assumptions concerning independence we use a heuristic iterative approach on the calibration samples to estimate the critical threshold. The critical 3σ limit is calculated by iteratively testing one reference sample against the remaining reference samples, creating a distribution of $z_{\max}$ values ($z_{\max,1}, z_{\max,2}, \ldots, z_{\max,25}$). From this a 3σ limit is calculated as:

$$Limit_{3\sigma} = \bar{z}_{\max} + 3sz_{\max} \quad (9)$$

where

$$\bar{z}_{\max} = \frac{1}{M} \sum_{i=1}^{M} z_{\max,i} \quad (10)$$

and

$$sz^2_{\max} = \frac{1}{M} \sum_{i=1}^{M} (z_{\max,i} - \bar{z}_{\max})^2 \quad (11)$$

M is the number of reference samples.

3. Experimental

Thirty in-process samples of the insulin intermediate DesB30 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 30 samples represent a substantial time period representing possible changes in production. One sample was spiked with human insulin drug product in five various levels from 0.01% to 0.15%. Human insulin is co-eluting with the structurally related target compound DesB30-insulin, but has a different molecular weight and thus different *m*/*z* values. Samples were injected into a gradient (0.05% TFA/10% acetonitrile and 0.05% TFA/70% acetonitrile) LC–MS system consisting of an Alliance reverse phase HPLC system (Waters, MA, USA), a Kinetex C18 column (150 mm × 3 mm, 2.6 μm) (Phenomenex, CA, USA), and a MicroTOF-Q II mass spectrometer (Bruker Daltonics, Bremen, D) operated with electrospray (ESI) in the positive ion mode. ESI provides maximum intentsity of the MH^{4+} ions, why this charge state was used in the calculations. All 30 NOC samples were measured in one replicate, whereas the five spiked samples were measured in five replicates each. The LC–MS data was collected and exported as text files using a software tool called DataAnalysis (Bruker Daltonics) and imported to Matlab version 7 (Mathworks, MA, USA) for further analysis. All software was written in Matlab using tools from PLS_Toolbox (Eigenvector Research, WA, USA) and Statistics Toolbox (Mathworks).

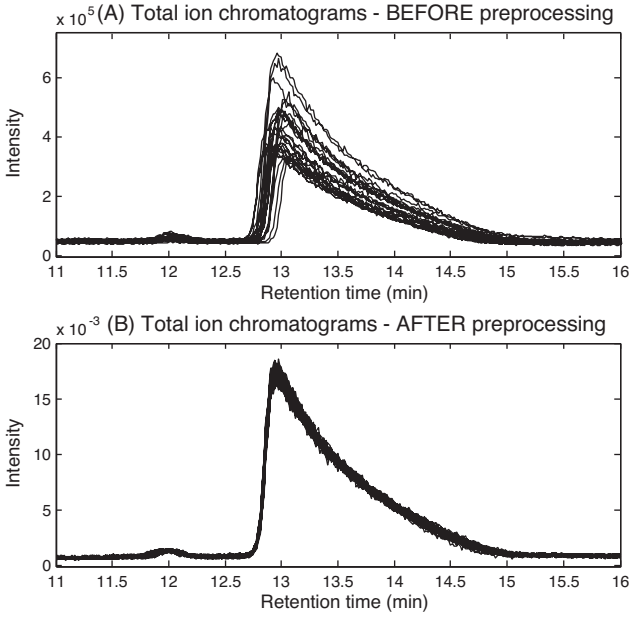

Fig. 3. TIC profiles of all samples before (A) and after (B) preprocessing (baseline correction, normalization and time alignment).

4. Results and discussion

4.1. Initial phase

The 55 LC–MS samples (30 NOC samples and 5×5 spiked samples) were collected and organized as an $M \times N \times O$ dataset *X*, with *M* samples, *N* elution times, and *O* *m*/*z* values. A relevant LC–MS window was chosen around the target peak, resulting in a 55 (samples) × 300 (retention times) × 200 (*m*/*z* values) dataset. For baseline correction of the data, a second order polynomial was fitted to each extracted ion chromatogram from each sample, based on 50% of all data points. The settings were chosen upon initial investigation of different alternatives. Once the samples were normalized by the sum of all intensities for each sample, time alignment using *i*coshift was sufficient for proper alignment of the LC–MS data. The corrected time axis was calculated from the total ion chromatogram (TIC) profiles, and then applied to the extracted ion chromatograms (EIC) of the corresponding LC–MS sample. The profile which showed the highest correlation with the remaining TIC profiles was selected as the target. For illustrative purpose, the corrected versus the original TIC profiles for all samples is presented in Fig. 3.

To make the measurements more comparable, and to reduce the huge amount of data points per sample, all intensities within a user-specified bin were summed. In this study, intensities within a bin size of 0.5 min and 2 *m*/*z* were summed. The bin size was chosen so that single peaks were approximately represented within in a bin. The binning reduced the number of data points from 60.000 to 1000 bin values per sample. Finally, a logarithmic transformation was used to adjust the variation in fold differences between the target peak and minor surrounding peaks, and to reduce the heteroscedasticity of the noise [33]. In Fig. 4, the effect of data reduction transformation of a NOC sample is illustrated, showing that smaller features around the target compound are enlarged due to binning and scaling.

4.2. Training phase

The essence of the training phase is to model the common cause variation present in the LC–MS samples obtained under normal operating conditions. The number of samples needed to construct a representative NOC model and control charts depends on the application. In this case study, a calibration set consisting of the first 25 chronologically ordered LC–MS NOC samples was used to develop a two component PCA model describing nearly 82% of the variation. The optimal number of PCA components to include in the model was selected based on the variance captured and on the results of leave-one-out cross-validation (data not shown). Variance captured flattens out somewhat after two components, and root mean squared error of cross-validation (RMSECV) has a clear local minimum at two components, indicating that after this point, the components just reflect noise. Furthermore, the inspection of loadings confirmed that the first two components reflect real chemical variation (Fig. 5).

The model was validated using an independent validation set consisting of the last five LC–MS samples. By inspection of the *D*- and *Q*-statistics (Fig. 6) it was confirmed that two components describe the common-cause variation. All 30 NOC samples were within the 95% quantile in both the *D*-statistic chart and the *Q*-statistic chart.

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 the other hand, when developing the model in the training phase, both the

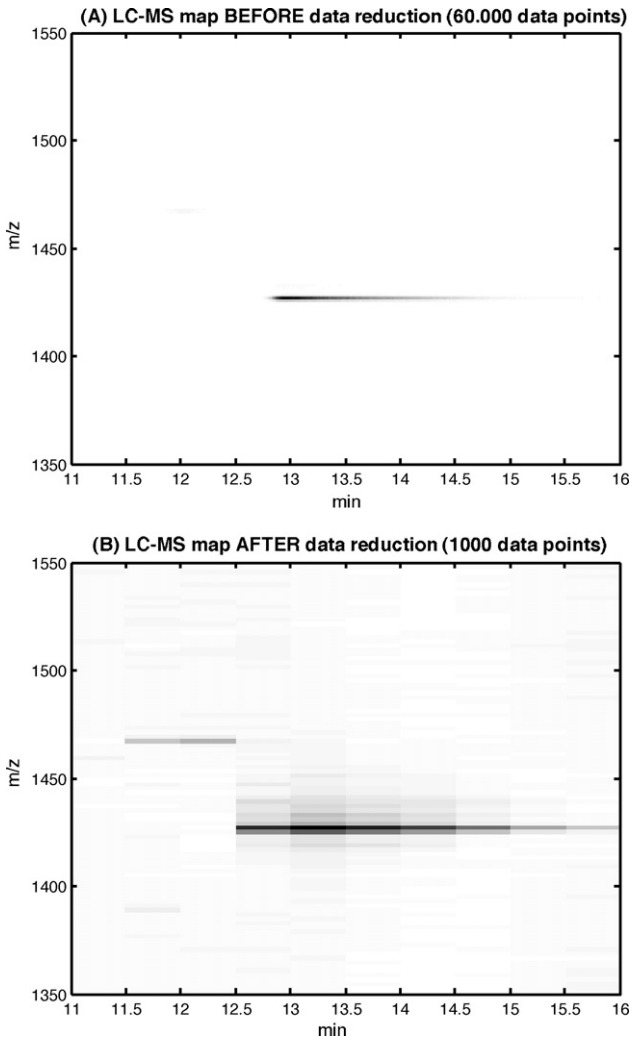

Fig. 4. LC–MS maps before (A) and after (B) data reduction of 60.000 data points to 1000 bin values, using a bin size of 0.5 min and 2 *m*/*z*.

D- and *Q*-statistics are of interest. These statistics may contribute with important and complementary indications about samples to exclude from the NOC model due to deviation in *common-cause variation* (*Q*) and *magnitude* (*D*). In this case all 30 samples used

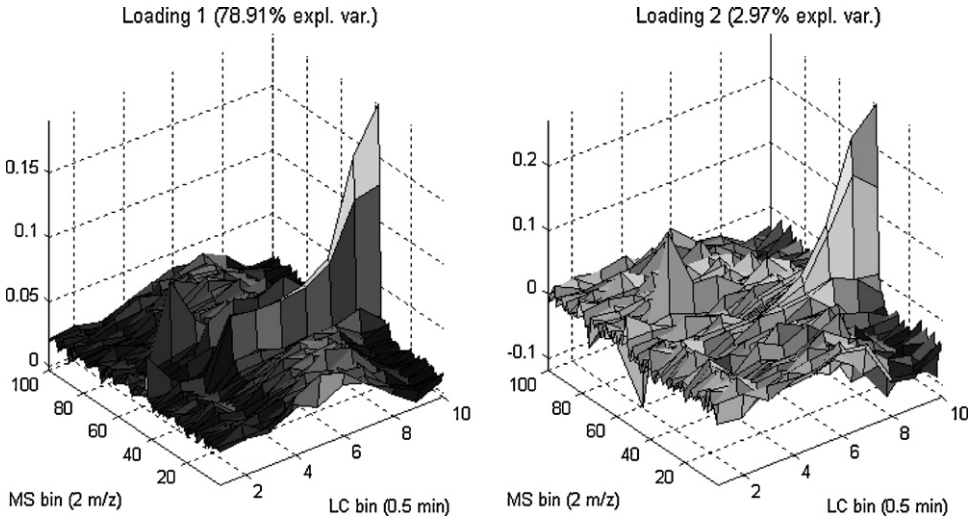

Fig. 5. 3D plot of the first two PCA loadings.

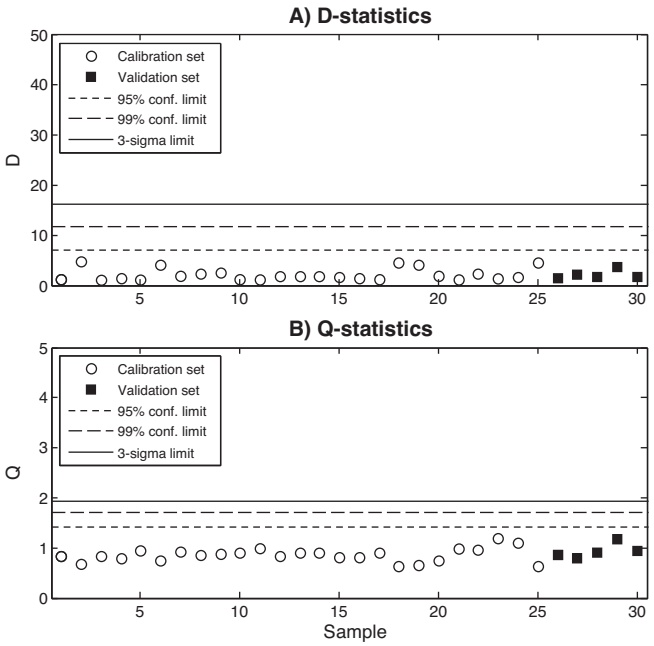

Fig. 6. Plot of (A) *D*-statistics and (B) *Q*-statistics of calibration (circle) and validation (square) sample sets.

in the training phase are within their respective 3σ limits in both *D*- and *Q*-statistics charts, and are therefore assumed to describe common-cause variation.

4.3. Application phase

To demonstrate the lack of sensitivity of ordinary *Q*-based MSPC applied to LC–MS data, a sample from the validation set was spiked with human insulin drug product in five various levels from 0.01% to 0.15%. Human insulin is co-eluting with the structurally related target compound DesB30-insulin, but has slightly different *m*/*z* values. The five spiked samples (measured in five replicates) were used to evaluate the ability of detecting an unknown impurity co-eluting with the target compound. As indicated in the *Q*-statistic chart (Fig. 7) none of the simulated chromatograms were detected as faulty by falling outside the 3σ limit.

As discussed earlier the *Q*-statistic measure suffers from lack of sensitivity due to the needle-in-the-haystack expedition. In Fig. 8 the *Q* contributions are presented for a sample spiked with 0.15% impurity. Even though the contributions provide indications of an abnormality around *m*/*z* 1450–1454 eluting at 12.5–13 min, the

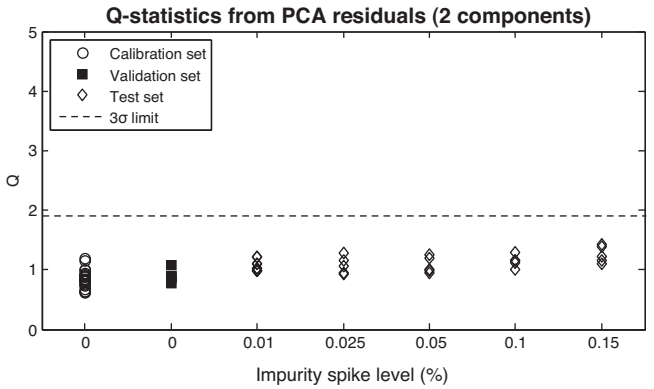

Fig. 7. Plot of *Q*-statistics of calibration- (circle), validation- (square), and test samples (diamond).

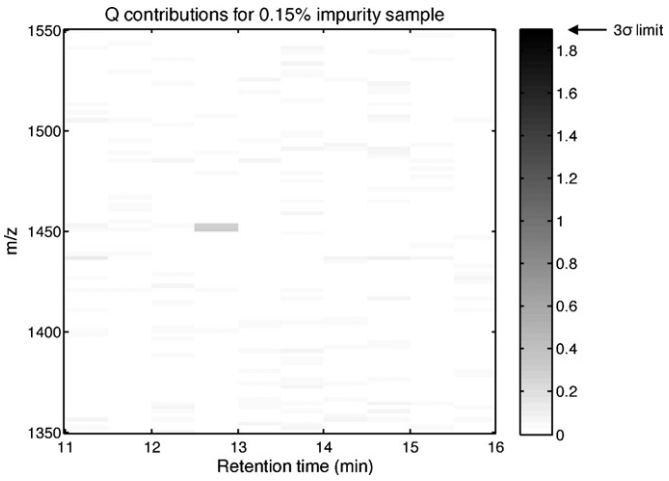

Fig. 8. Plot of *Q* contributions from PCA prediction of sample spiked with 0.15% HI.

relevant diagnostics seems to drown when calculating *Q*. As a consequence the relevant information is not detected and exploited.

Therefore, possible deviations were detected in the *individual* bins using multiple testing rather than testing the overall residual variation. The critical 3σ limit was calculated by iteratively testing one calibration sample against the remaining calibration samples. In comparison with the theoretically derived critical value (4.55), the data generated 3σ limit is slightly lower (3.75). This controversy is primarily due to the incorrect independence assumption which produces a more conservative limit, but maybe also deviation from the normality assumption in the *t*-tests. As indicated in Fig. 9, spike levels down to 0.05% HI was detected as faulty, falling outside the 3σ limit.

The detection level was tested using different selections of bin size and consequently bin number. The detection level is here defined as the lowest spike level where all five replicate samples were detected as faulty, falling outside the 3σ limit. In Fig. 10 the results of different selections of bin size and corresponding impurity detection level is presented. It appears from Fig. 10 that the lowest detection level is obtained with a bin size from 30 to 60 s and 1–2 *m*/*z* value. The number of bins in that region varies from 500 up to 2000 bins. Clearly too high complexity in terms of number of bins will result in a higher critical test limit followed by a higher level of detection. On the other hand in a coarse binning the signal disappears with higher level of detection as a consequence. Though the same consequence, the origin is different for the two cases. For high number of bins the detection limit is dependent on the false positive control in the modeling part, whereas for coarse binning effect

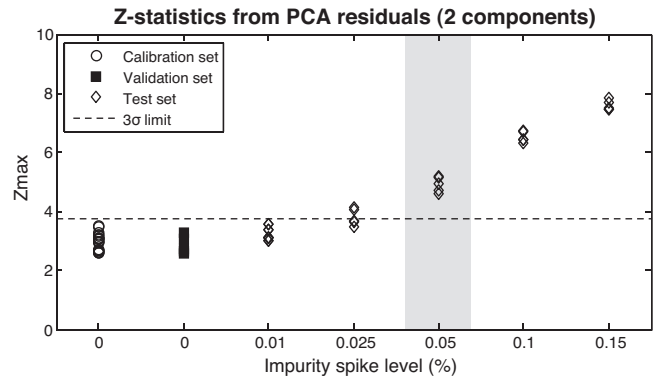

Fig. 9. Plot of *Z*-statistics of calibration- (circle), validation- (square), and test samples (diamond).

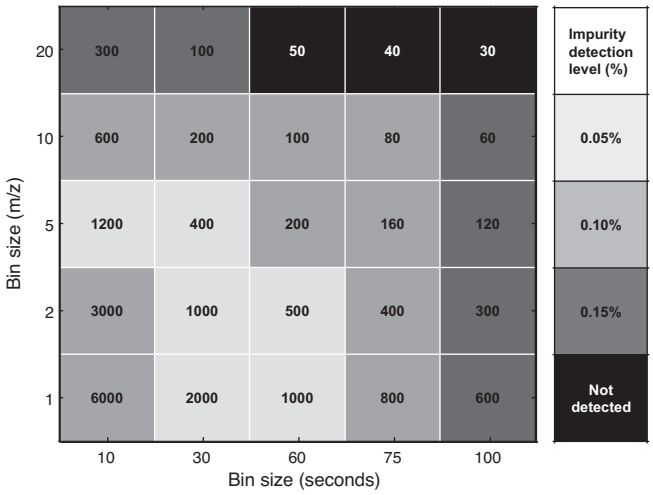

Fig. 10. Results of different selections of bin size (and number of bins) and corresponding detection level. The total bin numbers are indicated in the figure for each bin setting and colored according to the impurity detection level.

size vanishes for impure samples. Of course the impurity detection level examination presented here is optimized for this particular impurity, and is hence slightly biased downwards due to selection of bin size. For a true detection level determination an independent test set could be applied using the selected bin size. Ideally, a more objective method for selection of bin size should be considered. This is more likely an analytical discipline rather than a mathematical discipline. Future unknown impurities eluting close to the drug substance are most likely structurally related to the drug substance, and the impurities can be expected to show up in a 1000-fold difference compared to the drug substance. Hence, the examination presented in Fig. 10 may not be that misleading, and could serve as a preliminary bin-tuning procedure before setting up a reliable monitoring scheme.

To determine those variables responsible for the faulty detection the *Z* contribution plot is examined (Fig. 11). Clear diagnostics of the detected sample is provided, indicating that an unknown impurity is found around *m*/*z* 1450–1454 eluting around 12.5–13 min. Further inspection of the highlighted area (data not shown) revealed clear ion trace signals with a maximum intensity at *m*/*z* 1453. For more detailed diagnostics an extracted ion chromatogram (EIC) of *m*/*z* 1453 can be examined (Fig. 12).

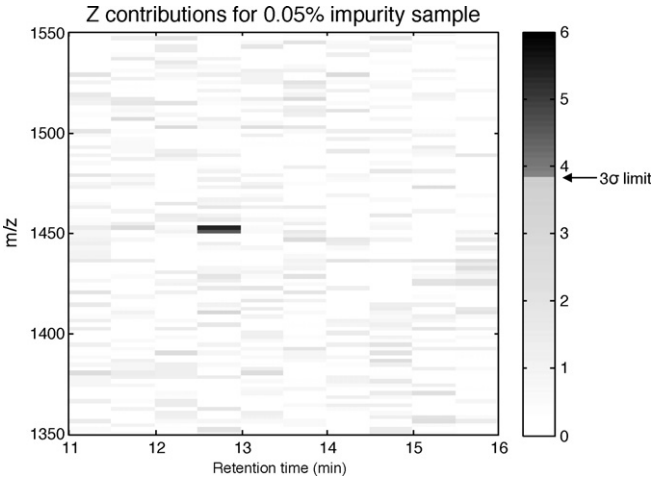

Fig. 11. Plot of *Z* contributions from PCA prediction of sample spiked with 0.05% HI.

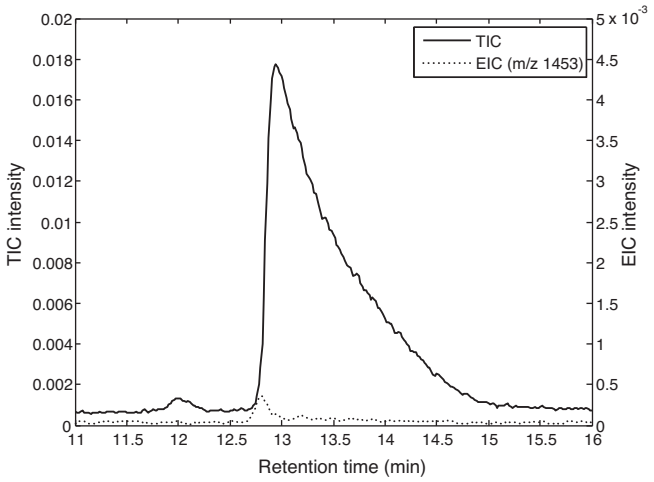

Fig. 12. Plot of TIC and EIC (*m*/*z* 1453) of sample spiked with 0.05% HI.

From the EIC the elution profile of the unknown impurity is provided. It would be difficult or impossible to detect a co-eluting 0.05% impurity peak if measured with HPLC. However with LC–MS this challenge is possible and becomes practicable if assisted by the automated methods demonstrated in this study. However, it is important to clarify that MSPC should not be regarded as a replacement of analytical knowledge when interpreting the LC–MS data. Instead, MSPC should be seen as the means for creating robust and highly interpretable multivariate models with the aim of monitoring and detecting unknown features in large and complex LC–MS data.

5. Conclusion and perspectives

This study demonstrates that MSPC based on PCA in conjunction with multiple testing is very powerful for monitoring and detection of unknown and co-eluting impurities measured with LC–MS. A spiked impurity present at low concentrations (0.05%) was detected and comprehensible contribution plot containing clear diagnostics of the unknown impurity was provided. From examination of contribution plots for lower spike levels than 0.05% (0.025% and 0.01%) large contributions from the unknown impurity were highlighted, emphasizing the sensitivity of this method. Trading off false negative signals by lowering of the critical limit from e.g. 3σ to 2σ might enhance the detection limit further. This tool will monitor and highlight only relevant areas in the complex LC–MS data where further effort on interpretation should be applied. Furthermore the tool proved robust towards treating instrumental artifacts such as baseline- and retention time drift. Applying this procedure for the detection of new peaks makes a fully automatic monitoring of LC–MS data possible. Furthermore, if implemented and operating while the purity analyses runs, this tool may considerably reduce time needed for subsequent assessment of data, and operate according to the PAT concept aiming for real-time release. Obviously the actual root cause of the alarm is not automatically given when applying this tool. Such an analysis would need incorporation of chemical and technical process knowledge and possibly applying MS/MS fragmentation for further compound identification. Label-free LC–MS data analysis is already widespread in proteomics and may well be increasingly important in the pharmaceutical industry. However, many different types of applications can be developed with LC–MS. Due to such variety of possible applications and approaches it may also be challenging to develop and incorporate a generic solution for processing and analysis of LC–MS data in commercial software. Nevertheless, this study

point towards development and incorporation of more advanced multivariate data analysis methods in commercial software solutions.

Acknowledgements

The authors thank Professor Rasmus Bro (University of Copenhagen, Department of Food Science), and Niels Væver Hartvig (Novo Nordisk A/S, Compliance Support) for their helpful suggestions on revising this paper.

References

[1] International Conference on Harmonization (ICH) – Guidance for Industry: Q3A(R2) Impurities in New Drug Substances, 2006.
[2] International Conference on Harmonization (ICH) – Guidance for Industry: Q3B(R2) Impurities in New Drug Products, 2006.
[3] K. Wiberg, M. Andersson, A. Hagman, S.P. Jacobsson, J. Chromatogr. A 1029 (2004) 13.
[4] C.K. Lim, G. Lord, Biol. Pharm. Bull. 25 (2002) 547.
[5] M.J. Frederiksson, P. Petersson, B.-O. Alexsson, D. Bylund, J. Sep. Sci. 32 (2009) 3906.
[6] D. Lincoln, A.F. Fell, N.H. Anderson, D. England, J. Pharm. Biomed. Anal. 10 (2010) 837.
[7] J.S. Salau, M. Honing, R. Tauler, D. Barceló, J. Chromatogr. A 795 (1998) 3.
[8] D. Bylund, R. Danielsson, K.E. Markides, J. Chromatogr. A 915 (2001) 43.
[9] E. Per-Trepat, S. Lacorte, R. Tauler, J. Chromatogr. A 1096 (2005) 111.
[10] K. Laursen, S.S. Frederiksen, C. Leuenhagen, R. Bro, J. Chromatogr. A 1217 (2010) 6503.
[11] H. Hotelling, J. Educ. Psychol. 24 (1933) 417.
[12] S. Wold, K. Esbensen, P. Geladi, Chemometr. Intell. Lab. Syst. 2 (1987) 37.
[13] P. Ralston, G. DePuy, J.H. Graham, ISA Trans. 43 (2004) 639.
[14] H.J. Ramaker, E.N.M. van Sprang, S.P. Gurden, J.A. Westerhuis, A.K. Smilde, J. Process Control 12 (2002) 569.
[15] J. Listgarten, A. Emili, Mol. Cell. Proteomics 4 (2005) 419.
[16] F. Gan, G. Ruan, J. Mo, Chemometr. Intell. Lab. Syst. 82 (2006) 59.
[17] S.J. Callister, R.C. Barry, J.N. Adkins, E.T. Johnson, W.J. Qian, B.J. Webb-Robertson, R.D. Smith, M.S. Lipton, J. Proteome Res. 5 (2006) 277.
[18] F. Savorani, G. Tomasi, S.B. Engelsen, J. Magn. Reson. 202 (2010) 190.
[19] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[20] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231.
[21] R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde, M. van der Werf, BMC Genomics 7 (2006) 142.
[22] W. Urfer, M. Grzegorczyk, K. Jung, Proteomics 6 (2007) 48.
[23] M.C. Wiener, J.R. Sachs, E.G. Deyanova, N.A. Yates, Anal. Chem. 76 (2004) 6085.
[24] J. Listgarten, R.M. Neal, S.T. Roweis, P. Wong, A. Emili, Bioinformatics 23 (2007) E198.
[25] S. Wold, Technometrics 20 (1978) 397.
[26] A. Ferrer, Qual. Eng. 19 (2007) 311.
[27] H. Hotelling, in: C. Eisenhart, M.W. Ha0stey, W.A. Wallis (Eds.), Techniques of Statistical Analysis, McGraw-Hill, New York, 1947, p. 113.
[28] J.E. Jackson, A User's Guide to Principal Components, John Wiley and Sons, 1991.
[29] J.E. Jackson, G.S. Mudholkar, Technometrics 21 (1979) 341.
[30] J.D. Storey, R. Tibshirani, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 9440.
[31] Y. Benjamini, Y. Hochberg, J. R. Stat. Soc. Ser. B: Methodol. 57 (1995) 289.
[32] J. Trygg, E. Holmes, T. Lundstedt, J. Proteome Res. 6 (2006) 469.
[33] O.M. Kvalheim, F. Brakstad, Y. Liang, Anal. Chem. 66 (1994) 43.