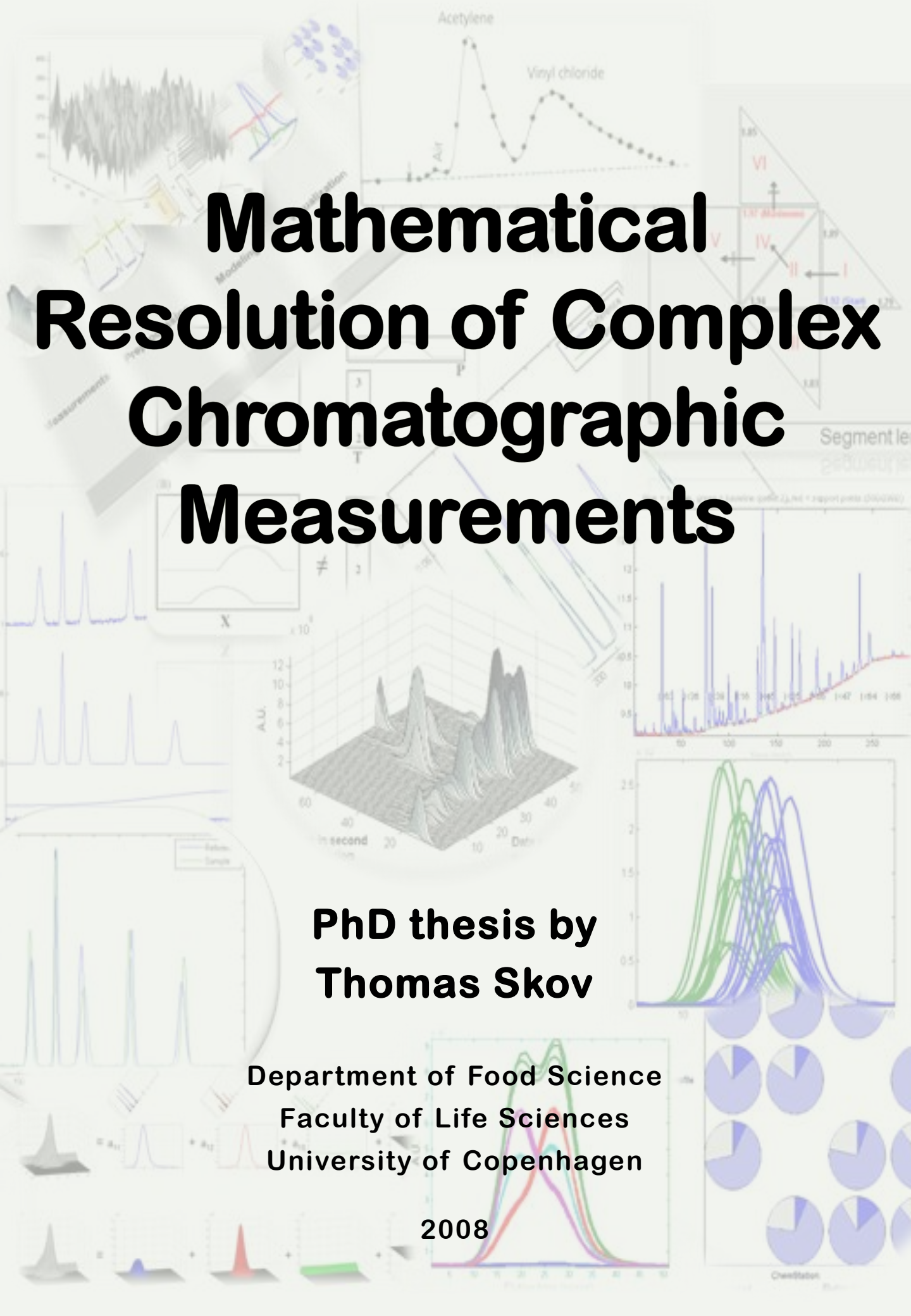


Mathematical Resolution of Complex Chromatographic Measurements

PhD thesis by
Thomas Skov

Department of Food Science
Faculty of Life Sciences
University of Copenhagen

2008

Preface

This PhD thesis has been conducted at the Quality and Technology section, Department of Food Science, Faculty of Life Sciences at the University of Copenhagen. The grant was provided by the FOOD Graduate School and financed partly by ARLA Foods amba and University of Copenhagen. I am thankful to ARLA Food amba for letting me be a part of their inspiring research environment.

This work has been supervised by Professor Rasmus Bro. He introduced me to the world of multiway chemometrics and has since then been a great source of inspiration. I am very grateful and appreciate the many challenges he has put me through over the last three years from teaching classes, writing papers, traveling the world, administrating courses and not forgetting buying the table tennis table! Looking back, I cannot find one single challenge that I would not accept again.

I would also like to thank Rob E. Synovec from University of Washington, Seattle, US, for letting me stay in his research group for almost three months during spring 2007. It was a special time for me that I hope to relive one day.

My time at Quality and Technology has been truly amazing. All the people here have brought so much laughter, good mood and *variability* into my daily life that I would not hesitate to describe them as one of my major "*principal components*" over the last three years. Special thanks go to Karin Kjeldahl and Peter Ibsen Hansen for our endless discussions of the small curiosities in life. I also thank Lars Nørgaard and Frans van den Berg for taking my knowledge in chemometrics to a new and still progressing level. And of course last but not least, thanks to all my coworkers, coauthors and collaborators to whom I will always be thankful.

The final appreciation goes to my family and friends – without them, my PhD study would definitely not have been the same!

A PhD is like chromatography – full of blurred ups and downs but eventually it peaks!

Thomas Skov
Copenhagen
15th of May 2008

Abstract

Mathematical resolution of complex chromatographic measurements has been a major issue since the invention of the first chromatographic separation instruments. Today the available columns are significantly better than decades ago. But the complexity of the samples analyzed and the need for faster chromatographic separations have diminished the effect of the increased resolution capabilities. Artifacts observed in the early days of chromatography are still appearing and must be dealt with to obtain data where the relevant information is readily accessible. These artifacts are mainly shifted peaks, background offset and co-elution of peaks. Several methods have been put forward to handle these artifacts, but they are often not accessible to laymen and only of interest for the scientific society focusing on the same areas (chemometricians).

This thesis puts forward and discusses selected pre-processing and processing (modeling) methods that have proven successful for chromatographic data of diverse structure and dimensionality. This includes 1) to automate and simplify the understanding of alignment and baseline correction methods including a comparison of the most popular ones for chromatographic data, 2) to visualize how rather complex mathematical approaches can extract and describe real chromatography (i.e. relevant information that would otherwise be hidden behind artifacts) and 3) to give chromatographers, not familiar with chemometrics (an overall description of the applied mathematical methods), insight into how these methods work. Consequently, all methods have been presented, explained and visualized on chromatographic data with more detailed information found in the attached PAPER I-V.

The first paper (PAPER I) describes a method to obtain an efficient alignment based on a simplex coordinate optimization routine. This is done calculating specific novel quantitative measures describing the alignment process. From this, the optimal or near-optimal alignment can be found with limited user-interaction and knowledge of alignment techniques. PAPER II, IV and V introduce multiway methods; PARAFAC and especially PARAFAC2 for different chromatographic data structures (GC-MS and GC×GC-TOFMS). The pros and cons of these methods are discussed with respect to selectivity, signal-to-noise ratio and degree of shift. PAPER III demonstrates a novel method to compare information (i.e. variation) from multiple data blocks. The available information is split into a unique, a common and a non-descriptive part and in this way redundancy and commonalities between multiple data blocks can be explored.

Resumé

At kunne adskille og beskrive toppe i komplekse kromatografiske målinger ved hjælp af matematik har været i søgelyset siden udviklingen af det første kromatografiske instrument. Undervejs er de tilgængelige kromatografiske kolonner blevet markant bedre. De prøver der skal analyseres er dog samtidigt blevet mere komplekse og da man ofte ønsker en adskillelse af de kemiske komponenter på kortere tid, så kan effekten af de forbedrede separationsegenskaber ikke udnyttes fuldt ud. Dette har betydet, at artefakter observeret i starten af kromatografiæraen stadig findes i data og at disse skal håndteres før den relevante information i data er tilgængelig. Disse artefakter er hovedsageligt forskudte toppe, basislinje- eller baggrundsbidrag samt overlappende toppe. Der er fremlagt flere metoder til håndtering af disse artefakter, men ofte er metoderne ikke tilgængelige for lægmænd og udelukkende af interesse for forskningsgrupper, der allerede fokuserer på disse områder (kemometrikere).

I denne ph.d. afhandling præsenteres og diskuteres udvalgte forbehandlings- og modelleringsteknikker, der har vist at være succesfulde for forskellige typer af kromatografiske data. Dette indbefatter 1) at automatisere og simplificere forståelsen af alignment (korrektion af forskudte toppe) og basislinjekorrektionsteknikker samt at inkludere en sammenligning af de mest populære metoder for kromatografiske data, 2) at visualisere hvordan komplekse matematiske metoder er i stand til at ekstrahere og beskrive reelle kromatografiske data (dvs. beskrive relevant information, som ellers ville være skjult bag de nævnte artefakter) og 3) at give kromatografispecialister, der ikke er familiære med kemometri (som er en samlet betegnelse for de anvendte matematiske metoder), et indblik i metodernes virkemåde. Derfor er alle metoderne beskrevet, forklaret og visualiseret ud fra kromatografiske data, mens mere detaljeret information kan findes i de enkelte vedlagte artikler (PAPER I-V).

Den første artikel (PAPER I) beskriver en metode, der baseret på en optimeringsrutine kan finde frem til en ideel alignment af kromatografiske toppe. Dette gøres ved at optimere kvantitative mål, der beskriver alignmentprocessen. Ud fra dette kan den optimale eller nær-optimale alignment findes med minimal brugerinteraktion og viden om alignmentteknikker. PAPER II, IV og V introducerer multivejs metoder; PARAFAC og PARAFAC2 for forskellige kromatografiske data (GC-MS og GC×GC-TOFMS). Fordele og ulemper ved disse metoder bliver diskuteret i relation til selektivitet, signal-støj forhold og graden af forskudte toppe. PAPER III demonstrerer en ny metode, der sammenligner information (dvs. variation) fra flere datablokke. Dette gøres ved at splitte den tilgængelige information op i en unik, en fælles samt en ikke-forklarende del. Ved dette opnås et overblik over den redundante og den fælles variation, der findes mellem datablokkene.

List of publications

PAPER I

Skov, T., van den Berg, F., Tomasi, G., and Bro, R. **2006**. Automated alignment of chromatographic data. *Journal of Chemometrics*, 20 (11-12): 484-497.

PAPER II

Skov, T. and Bro, R. **2008**. Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry*, 390 (1): 281-285.

PAPER III

Skov, T., Ballabio, D., and Bro, R. **2008**. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Analytica Chimica Acta*, 615 (1): 18-29.

PAPER IV

Amigo, J.M., Skov, T., Coello, J., Maspoch, J., and Bro, R. **2008**. Solving GC-MS problems with PARAFAC2. *Accepted for publication in Trends in Analytical Chemistry - TrAC*.

PAPER V

Skov, T., Hoggard, J.C., Bro, R., and Synovec, R.E. **2008**. Handling Retention Time Shifts in GC×GC-TOFMS Data using Shift Correction and Modeling. *In preparation*.

Additional paper by the author

Skov, T. and Bro, R. **2005**. A new approach for modelling sensor based data. *Sensors and Actuators B – Chemical*, 106 (2): 719-729.

List of abbreviations

ALS	Alternating Least Squares
AMDIS	Automated Mass spectral Deconvolution and Identification System
COW	Correlation Optimized Warping
EFA	Evolving Factor Analysis
FID	Flame Ionization Detector
GC	Gas Chromatography
GC×GC	Comprehensive two-dimensional Gas Chromatography
HPLC	High Performance Liquid Chromatography
IR	InfraRed spectroscopy
LC	Liquid Chromatography
MCR	Multivariate Curve Resolution
MS	Mass Spectrometry
MVP	Multiblock Variance Partitioning
NIR	Near InfraRed spectroscopy
PARAFAC	PARAllel FACtor analysis
PARAFAC2	PARAllel FACtor analysis 2
PCA	Principal Component Analysis
PLS(R)	Partial Least Squares (Regression)
PTW	Parametric Time Warping
PWA	Piecewise Alignment
SIC	Single Ion Chromatogram
SIM	Selected Ion Monitoring
STW	Semi-parametric Time Warping
SVD	Singular Value Decomposition
TIC	Total Ion Count
TOFMS	Time of Flight mass analyzer
UV	Ultra Violet spectroscopy

The notations used for **alignment**, **baseline**, **factor models** etc. are mentioned in the respective chapters.

Contents

PREFACE ..... **I**

ABSTRACT ..... **II**

RESUMÉ ..... **III**

LIST OF PUBLICATIONS ..... **IV**

LIST OF ABBREVIATIONS ..... **V**

CONTENTS ..... **VI**

1 INTRODUCTION **1**

1.1 Aim of thesis **2**

1.2 Thesis outline **3**

2 CHROMATOGRAPHY – *SETTING THE SCENE* **5**

2.1 Chromatography – some historical aspects **5**

2.2 Parts of the chromatographic signal **7**

2.3 Dimensionality and structure of chromatographic data **9**

2.4 Chromatographic data processing **10**

2.4.1 De-noising and smoothing 11

2.4.2 Alignment 13

2.4.3 Baseline correction 15

2.5 Peak co-elution/overlapping peaks **16**

3 CHROMATOGRAPHIC DATA **21**

3.1 GC-MS **21**

3.2 GC×GC-MS **24**

4 PRE-PROCESSING OF CHROMATOGRAPHIC DATA **27**

4.1 Alignment **29**

4.1.1 The warping function/path 31

4.1.2 Alignment techniques used for fingerprint chromatographic data 33

4.1.3 Optimization of the alignment parameters 44

4.1.4 Further improvement of alignment 49

4.1.5 Alignment techniques used for GC-MS landscapes 50

4.2 Baseline correction 53
4.2.1 Methods for fingerprint chromatographic data 54
4.2.2 Methods for multidimensional chromatographic data 57
4.3 Peak finding 58

5 MODELING OF CHROMATOGRAPHIC DATA 61
5.1 Multivariate data analysis: Two-way 63
5.1.1 Single peak system – one sample (landscape) 66
5.1.2 Local fingerprints – more samples 67
5.1.3 Global fingerprints – more samples 69
5.2 Multivariate data analysis: Multi-way 70
5.2.1 PARAFAC 73
5.2.2 PARAFAC2 76
5.2.3 Uniqueness – *in short* 77
5.2.4 Determining the proper number of factors/components 81
5.2.5 Advanced methods vs. commercial chromatographic software 83
5.3 Advantages and thoughts when using advanced multivariate models 84

6 INFORMATION IN MULTIPLE DATA BLOCKS 93

7 CONCLUSIONS AND PERSPECTIVES 101
7.1 Pre-processing of chromatographic data 101
7.2 Modeling of chromatographic data 102
7.3 Information in multiple data blocks 103
7.4 Outro 103

8 REFERENCES 105

PAPER I-V 115

1 Introduction

The use of a chromatographic separation system involves several steps; preparing the sample, injecting it into the chromatographic column, separating the analytes, detecting the analytes, identifying the analytes followed by subsequent data analysis. Ideally the digitized signal (data) of the injected sample should contain the information needed to characterize the sample, to discriminate between samples or to classify samples into groups. However, this information is often hidden behind well known instrumentally induced artifacts[1].

Artifacts in chromatographic signals have been studied for several decades and novel and advanced algorithms have been applied to handle this. For chromatographic data the artifacts have essentially been of the same character since the invention of the first chromatographic systems (Ettre, 2008; James and Martin, 1952), but with the more advanced analytical instruments the need for modifications, novel and dedicated algorithms has increased significantly. The artifacts that are most significant in chromatographic signals and will be treated in this thesis are; peak shifts, baseline offset and overlapping peaks.

Chromatographic data are not the same as informative data – properly pre-processed and analyzed chromatographic data are as close to informative data as we can get![2]

The recent advances in instrumental technologies have enabled collection of vast amounts of data in a variety of scientific disciplines. Many of these datasets are described as being multidimensional, heterogeneous (complex), noisy, including artifacts and having a large degree of redundancy (correlated features) etc. For chromatographic data this is as advanced hyphenated techniques where eluting fractions are detected with multivariate systems (e.g. UV, MS). Often traditional techniques cannot be applied to these data sets either due to their massive size or due to their unconventional nature. Thus, novel algorithms have been designed to address the limitations of the existing techniques in handling the *challenges* created by these types of data [PAPER II].

Addressing these *challenges* requires close interaction between data mining algorithm designers (chemometricians) and application experts (chromatographers) since deep knowledge of the domain is the key to identifying and validating useful high-level features as well as determining patterns that are meaningful. The major challenge is to make sure that the chromatographers and the chemometricians speak the same language when creating and implementing new algorithms. The chromatographer might explain one topic based on terms

[1]Here artifacts will be defined as changes to the ideal chromatographic signal that either destroy or complicate the analysis of data.

[2]Text lines in boxes in bold and italic letters are statements put forward by the author. Citations are shown in italic letters with the corresponding reference.

from separation theory (column bleed, elution time, peak width etc.) where the chemometrician would use numerical or data based terms (baseline, scan number, number of data points per peak etc.).

Often chromatographers are limited to the data analysis tools available in the commercial chromatographic software package and they can only treat data in a certain way no matter the complexity of the samples, the degree of influential artifacts or the noise level. This means that the data analysis is often done with a lot of assumptions to the data structure, which is often not valid for proper data analysis. The advantage is that well-known methods are used and simple and understandable visualizations are available. The limitations are removed when exporting to numerical software that offers advanced and dedicated algorithms to solve all specific problems found in chromatographic data. But more advanced means more knowledge required and new ways of visualizing things. The aspects of getting chromatographers and chemometricians together were addressed recently by Daszykowski and Walczak (2006).

Chromatographers and chemometricians can be joined through more automated algorithms, improved visualization and an increased insight into both domains! In the end, this can lead to more powerful and advanced methods becoming useful to a broader range of users.

1.1 Aim of thesis

This thesis focuses on mathematical resolution of complex chromatographic measurements taking into account the many aspects present when analyzing chromatographic data (Figure 1). The intention of this thesis is not to develop and present new algorithms for pre-processing or for data modeling, but to modify and use current algorithms to enhance the collaboration between the chromatographic and chemometric society.

Focus has also been on automating and optimizing existing algorithms to limit the user-interaction and to present the results with simple and interpretable visualizations. Within this, methods for exporting (and handling) data from commercial chromatographic instruments to numerical software packages (e.g. MATLAB) to test and use the advanced algorithms have been pursued. Focus has also been on how to evaluate the extracted information from multiple analytical experiments that might contain redundant and unique information.

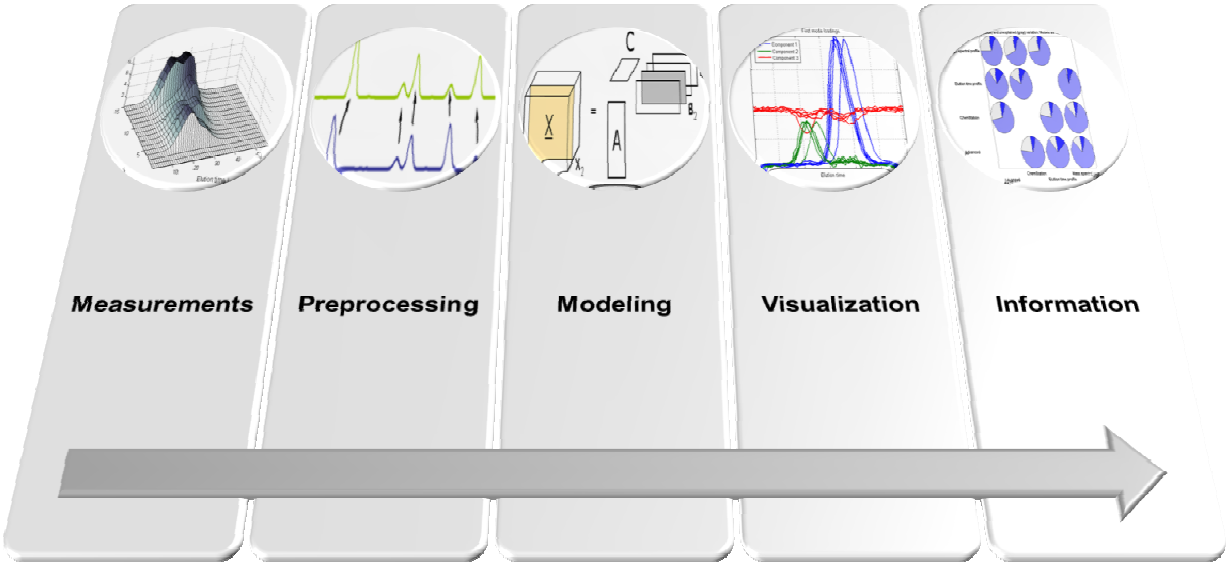


Figure 1. Aim of this thesis; from measuring the sample, preparing (pre-processing) it for the model step (modeling), providing a simple and interpretable visualization and ending up with a description of the available information.

1.2 Thesis outline

Chapter 2 – Chromatography – setting the scene
This chapter gives an introduction to the many chromatographic aspects and will set the scene for the main parts of the thesis.

Chapter 3 – Chromatographic data
This chapter gives a short introduction to chromatography in general as an analytical method. The instrumental properties such as hyphenation, detectors and primarily data structure and dimensionality are discussed.

Chapter 4 – Pre-processing of chromatographic data
Chapter 4 deals with pre-processing methods that can be used to prepare raw chromatographic data to the subsequent chemometric modeling. Especially, the focus will be on peak alignment and how methods for this can be made understandable to and used by a broader audience than just chemometricians. Also baseline correction will be considered and explained.

Chapter 5 – Modeling of chromatographic data
Analysis of chromatographic data often starts with a visual inspection of the data (e.g. by chromatographers) and ends with a visual inspection of the processed data (e.g. by chemometricians). The many steps from A to B involve the use of advanced algorithms and models. The applied model depends of the data structure available and as such different approaches can be used for chromatographic data. The data structure often depends on the expected outcome of the model; classification from fingerprints, peak integration from single or overlapping peaks and identification of the analytes in the sample. This chapter highlights aspects that must be considered when doing proper modeling of hyphenated chromatographic data.

Chapter 6 – Information in multiple data blocks
The last part of the thesis discusses how to use the information derived from chromatographic data. Different data representations are available for all chromatographic measurements depending on what should be extracted from or done with the data; peak areas, fingerprinting, classification. These representations give multiple blocks containing different kinds of information; some parts might be unique, some parts common and some parts might not provide any information about a given feature.

Chapter 7 – Conclusions and perspectives
A summary will be given of the major findings for chromatographic data analysis from data export to numerical software over the use of advanced chemometric methods to how results can be presented. Also, the topics where additional work is needed will be discussed.

2 Chromatography – *setting the scene*

In the early 1900s M.S. Tswett, the inventor of chromatography, stated that "*An essential condition for all fruitful research is to have at one's disposal a satisfactory technique*[3]" and he continued "*a scientist always must consider the whole sample and separate all the substances present*[4]".

These statements were and are indeed still valid for the chromatographic as well as other scientific disciplines. In the early years of chromatography the prevailing methodology in the study of complex mixtures was the *isolation* of a single chemical constituent and its purification using e.g. extraction and crystallization. In contrast, Tswett emphasized to look at the *whole* sample at a time and to separate all individual substances present from both the matrix and from one another. This change was fundamental and contrary to the belief of most of his contemporaries, but this statement has been more and more evident throughout the 20th century.

Also in the early 1900s the term 'technique' was more directed towards the instrumental part of the total chromatographic technique. In recent years it has been directed more and more towards the advanced and dedicated numerical techniques available. However, one cannot succeed without the other and today especially the chromatographic techniques prevail from advances in dedicated numerical solutions.

This chapter highlights the aspects of chromatographic data that, despite many years of significant improvement of the instrumental methods, still require subsequent numerical pretreatment to be able to extract the information in data.

2.1 Chromatography – some historical aspects

The word chromatogram comes from two Greek words meaning color and writing, which together, literally means color writing. The term originates from a study in the early 1900s by Tswett, a Russian botanist, in his separation of plant pigments [Tswett, 1906]. Tswett used a liquid-adsorption column containing calcium carbonate to separate plant pigments. The different pigments were separated along the column and gave a series of colored bands, which Tswett called a chromatogram. From the same terminology, the actual separation technique became known as chromatography. This was the start of liquid chromatography (later this technique was denoted adsorption chromatography), but it was not until 1941 that

[3]M.S. Tswett, Khromofilly v Rastitel'nom i Zhivotnom Mire (Chromophylls in the Plant and Animal Kingdom) (Karbasnikov Publishers, Warsaw, 1910). Cited from Ettre (2000).
[4]M.S. Tswett, Khromofilly v Rastitel'nom i Zhivotnom Mire (Chromophylls in the Plant and Animal Kingdom) (Karbasnikov Publishers, Warsaw, 1910). Cited from Ettre (2003).

the concept of gas chromatography was mentioned for the first time (Martin and Synge, 1941) and even later demonstrated how it worked in practice in a dissertation of a PhD by F. Prior in 1947 in Innsbruck, Austria [Ettre, 2008].

Today a chromatogram has virtually no association with color and although major changes to how the chromatographic signal is obtained have seen the light over the last 50 years, the visual appearance is virtually the same; a number of chemical constituents separated, detected and visualized as eluting peaks over time. The very first chromatogram of this type published in a scientific journal is hard to find, but printed chromatograms started to appear in the 1940s. An example of this is visualized in Figure 2.

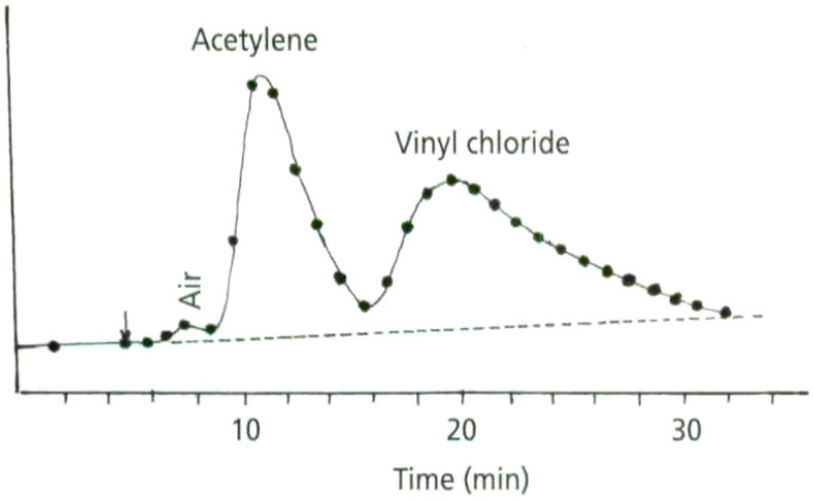


Figure 2. Separation of acetylene and vinyl chloride. Arrow: sample introduction (modified from Ettre, 2008). The detector was based on thermal conductivity [Bobleter, 1996]. Chromatogram presented in a PhD dissertation by F. Prior in 1947 in Innsbruck, Austria [Ettre, 2008].

The peaks in the chromatogram look fairly primitive due to the low number of data points per peak, but most importantly the peaks look similar to what can be found using today's more advanced chromatographic instruments. Artifacts such as broad (several minutes) and asymmetrical (tailing) peaks, drifting baseline and incomplete resolution were and are still topics that to some degree have to be solved. Smolkova-Keulemansova (2000) also mentioned that the first chromatogram obtained in 1946 was analytically uninteresting due to the, at that time, major problem of a drifting baseline.

Chromatograms of this form were also described and visualized by James and Martin (1952) in the early 1950s (example given in Figure 3) using a fairly complicated automatic titrator, which was applicable only to certain specific samples (e.g., free fatty acids or amines). In 1954 James and Martin, like F. Prior, introduced a thermal conductivity detector (similar to the one presented by Prior in 1947) using a milliammeter (measures the flow of an electrical current), which was further improved, described and visualized with chromatograms by Ray

[Ray, 1954a; Ray, 1954b]. This was the start of a fast evolution in GC detectors and soon the thermal conductivity detector was replaced by a potentiometric recorder. Today more advanced detectors are used to create the chromatograms, which are now presented as digitized signals.

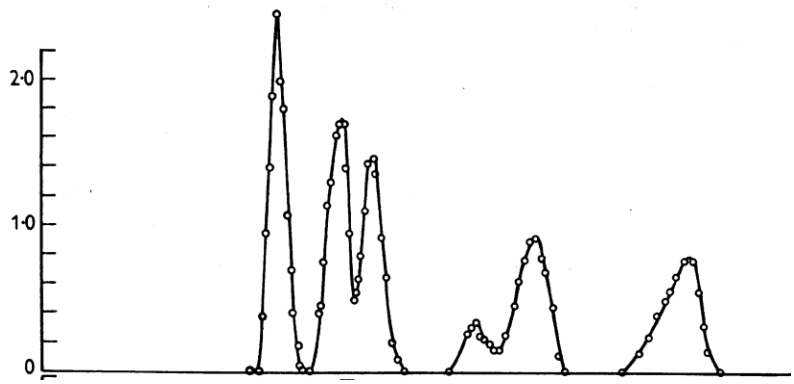


Figure 3. Separation of the isomers of valeric acid from n-butyric and isobutyric acids, showing incomplete resolution of several peaks (modified from James and Martin, 1952).

In Figure 3 several peaks overlap and at that time the solution to this problem was to construct a longer column. A longer column means more theoretical plates and higher column efficiency and an increase in the efficacy of the separation process (James and Martin, 1952), but at the expense of a longer chromatographic run.

To end this brief historical flashback on chromatography Hinshaw (2004) is quoted.

'*And yet, despite all the technology advances, chromatograms take nearly the same form today as they did 40 or 50 years ago: a signal dispersed in time and, in the case of multichannel detectors such as mass spectrometric or diode array types, multiple scans captured in sequence. Although they can occur quite rapidly, peaks still exist as deviations from a more or less noisy background, they have definable starting and stopping points, and, of course, chromatographers still need to find and measure them'*.

2.2 Parts of the chromatographic signal

Considering the univariate signal measured from e.g. an FID detector, the univariate signal measured from e.g. one mass channel from a multivariate detector or one elution time in one GC dimension in multidimensional GC, the intensity profile (overall signal) of a chromatogram can be divided into three constituting parts (Figure 4):

1) The analytical relevant signal
2) The background/baseline[5]
3) The noise

Summing these three parts provides the overall chromatographic signal that is the one provided by the chromatographic instrument.

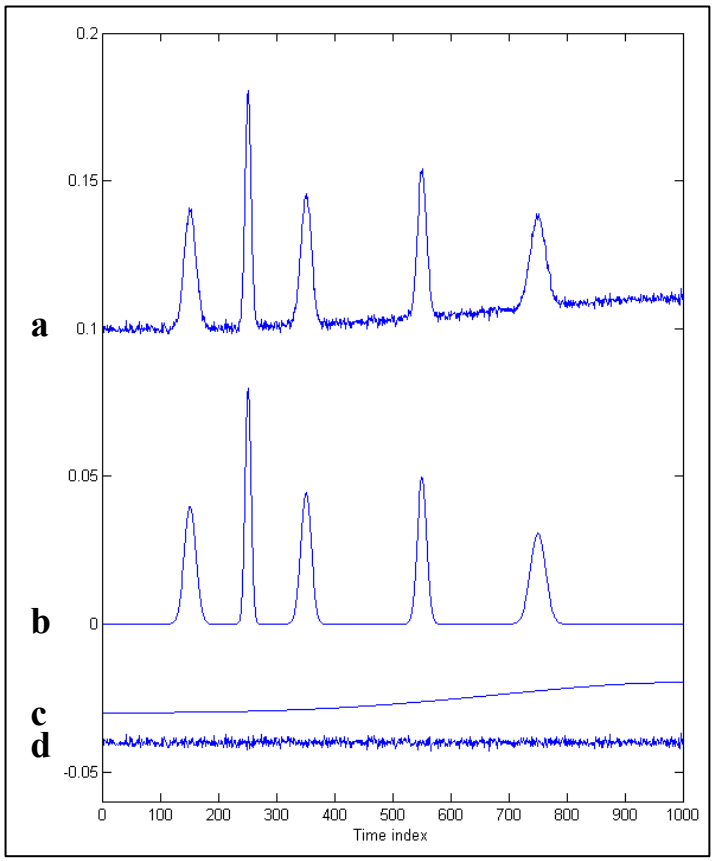


Figure 4. Components of the chromatographic analytical signal: (a) overall signal; (b) analytical relevant signal; (c) background/baseline; and, (d) noise (visualization inspired by Daszykowski and Walczak, 2006).

From the chromatographic signal several issues must be considered prior to the subsequent chemometric data analysis. For chromatographic data the major issues are baseline correction and peak alignment and to some degree also signal enhancement. The solution to these aspects is not straightforward and depends on one or more parameters to be set; parameters that depend on the data and problem at hand. These aspects are briefly presented in the following sections and further elaborated on in Chapter 4.

[5]The terms baseline and background will be used throughout this thesis for the same phenomena. However, when talking about the contribution to the analytical signal the term background is more appropriate. Dealing with the correction of the offset (e.g. as shown in Figure 4) the term baseline correction is more often used in the literature.

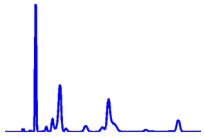
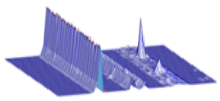
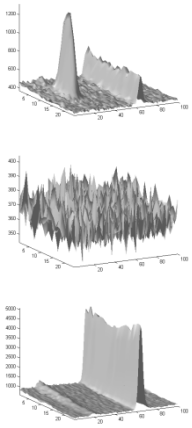
2.3 Dimensionality and structure of chromatographic data

The chromatographic instrumental and detector set-up provide data of different dimensions and with that different challenges and possibilities in the subsequent data analysis. The most important chromatographic and detector set-ups are listed below and the ones marked in bold will be further described and discussed in Chapter 3. In Table 1 different chromatographic instrumental systems are characterized with respect to e.g. dimensionality and terminology.

1) Chromatographic column(s)
    a. **GC**
    b. **GC×GC**
    c. GC×GC×GC or GC^3

2) Detector
    a. Univariate detectors
        i. Flame Ionization Detector (FID)
        ii. Thermal Conductivity Detector (TCD)
    b. Multivariate detectors
        i. **Quadrupole Mass Selective detector (MS)**
        ii. **Time of Flight detector (TOFMS)**

Table 1. Characterization of gas chromatographic data using different combinations of chromatographic set-up and detector.

Chromatography	**GC**	**GC**	**GC×GC**
Detector	FID	MS	TOFMS[1]
Terminology	GC-FID	GC-MS	GC×GC-TOFMS
Example			
Dimensions	Elution time[2]	Elution time × Mass channel	Elution time GC 1 × Elution time GC 2 × Mass channel
Size of one sample	Vector ($N = 1$)	Matrix ($N = 2$)	Three-way array ($N = 3$)
Size of dataset	Matrix	Three-way	(N+1)-way
Data properties	Bi-linear	Tri-linear	Quadri -linear[3]

[1]One mass channel per landscape for better visualization
[2]Instead of elution time the sample vector could also hold the integrated peak areas.
[3]In quadri-linear chromatographic data it is common to analyze one sample at a time and reduce the dimensionality to tri-linear data. This also removes the potential problem of shift in one of the GC dimensions across the samples.

2.4 Chromatographic data processing

Three broad categories can be set up that covers the majority of applications for chromatographic data analysis.

1) For data with perfect resolution and valid identifications for all peaks, little data processing is needed.
   A: Peak system[6]: find peak areas and identify peaks

[6]A *peak system* is defined as a local region of the chromatographic data with only one or few chemical compounds present/eluting. With a peak system, each sample is described by a two-dimensional matrix. E.g. two peaks in one sample (two-way GC-MS) or two peaks in several samples (three-way GC-MS). See also Chapter 5.

Typically chromatographic software packages (e.g. ChemStation by Agilent, 2001) handle this efficiently. This approach can be rather automated as long as a standard method including all target peaks (target peaks are peaks of interest and does not include very noisy and low intense peaks) is prepared.

B: Chromatographic fingerprints[7]

Chromatograms must be aligned. Subsequently, fingerprints are modeled with multivariate methods. Important peak regions can be identified and their characteristics further investigated if peaks have been identified from A.

2) If the resolution is poor, advanced curve resolution methods can be applied to separate co-eluted peaks into individual contributions (both elution time and mass spectral wise). Similar steps as in 1) but here it may be important to locate the problematic peak regions either prior to or after the initial multivariate modeling. Otherwise peak areas or fingerprints can be influenced by this.
3) With severe artifacts in the chromatograms such as baseline off-set, peak shifts and overlapping peaks, special attention to removing and handling these issues must be taken prior to multivariate modeling. Thus, focus must be on the fingerprints from the start as these artifacts will hinder proper peak finding, identification and integration.

There is an increasing tendency to use the entire chromatographic signal instead of using only the integrated peaks of interest. The chromatographic profile is then considered as a fingerprint that contains features unique to a given sample. The idea is to include as much information as possible instead of setting criteria for where in the chromatogram the information is placed. The latter is a prerequisite when using only the integrated peaks, as peaks falling outside certain criteria will be neglected. These criteria could be; 1) intensity above a certain threshold, 2) number of data points across the peak, 3) the signal-to-noise ratio and/or 4) the peak shape. This if further discussed in Chapter 5.

2.4.1 De-noising and smoothing

In a chromatographic signal, signal enhancement techniques such as noise filtering, smoothing and other techniques are performed to increase the signal-to-noise ratio for the peaks of interest and to remove the noise. This may confirm the presence of a peak in a noisy baseline, but it will not improve the precision or accuracy of e.g. peak integration or peak parameter estimation. Although signal enhancement techniques are efficient one should always consider the selected experimental conditions as a first shot to eliminate excessive noise.

[7] A *fingerprint* is defined as a local or global region of chromatographic data where a certain pattern or profile must be available for each sample. With a fingerprint each sample is described by a one-dimensional vector. E.g. collapsed mass spectral dimension – TIC. See also Chapter 5.

Several signal enhancement techniques are available and they can be divided into two broad classes; 1) digital filtering in the frequency domain or 2) in time domain. Digital filtering in the frequency domain separates noise from signal by dividing the overall signal into high frequency components and low frequency components. As the noise has a much higher frequency than the signal (Figure 4) the use of a low-pass filter allows the separation of noise and signal. In the time domain the filters usually work within a finite-width moving window of the chromatogram. The most widely used technique in analytical chemistry is the polynomial filter suggested by Savitzky and Golay [Savitzky and Golay, 1964; Steinier et al., 1972], which uses a least squares fit of a polynomial of a given order to a certain window size in the chromatogram. The polynomial evaluated at the center point is then used as the smoothed data and by moving the window from one end of the chromatogram to the other in successive steps the noisy data elements are replaced by fitted polynomial elements. The Savitzky and Golay smoothing method requires two input parameters; the window width and the order of the polynomial. These parameters will result in removal of either only a moderate degree of noise (Figure 5b) – high frequency noise, of an adequate degree of noise (Figure 5c) or of too much noise (of low frequency) that can filter out part of the chromatographic signal as well. Thus, the method requires some user experience before the optimal signal enhancement has been found.

As will be discussed in Chapter 5 the use of multivariate factor models can also increase the signal-to-noise ratio by separating signal from noise in different factor contributions.

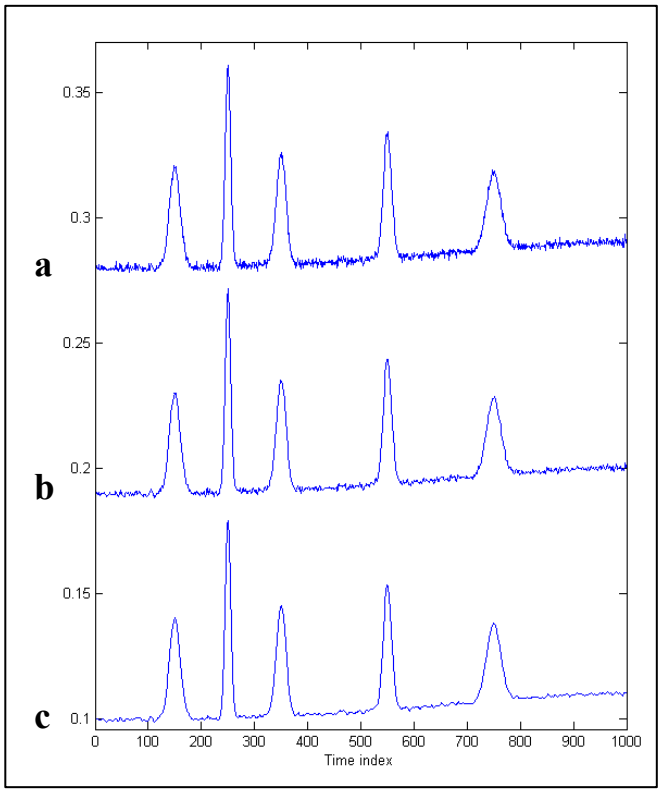


Figure 5. Signal enhancement visualization: (a) overall signal as indicated in Figure 4; (b) smoothed signal using Savitzky and Golay smoothing with a window width of 5 and a second order polynomial; (c) same as (b) but with a window width of 11.

2.4.2 Alignment

Elution time variations are often observed in chromatographic data due to subtle, random, and often unavoidable variations in instrument parameters. Pressure, temperature and flow rate fluctuations may cause an analyte to elute at a different elution time in replicate runs. Matrix effects and stationary phase decomposition may also cause elution time shifting. It is necessary to synchronize or align chromatographic data in the time dimension so that that the same peak is located at the same elution time for all samples. Analyzing individual samples can be achieved by using a proper indexing system, but when the samples are stacked to give higher order arrays, this is no longer a trivial problem.

In data matrices the same phenomenon or variation must be located in the same column for all samples to be able to conduct a meaningful analysis of data. This is a fundamental prerequisite when linear dependent methods such as classical statistics, pattern recognition techniques and the majority of advanced multiway methods are used. For chromatographic fingerprints, the deviation can be illustrated using a one peak approach, but the situation is the same for multiple peak systems (Figure 6).

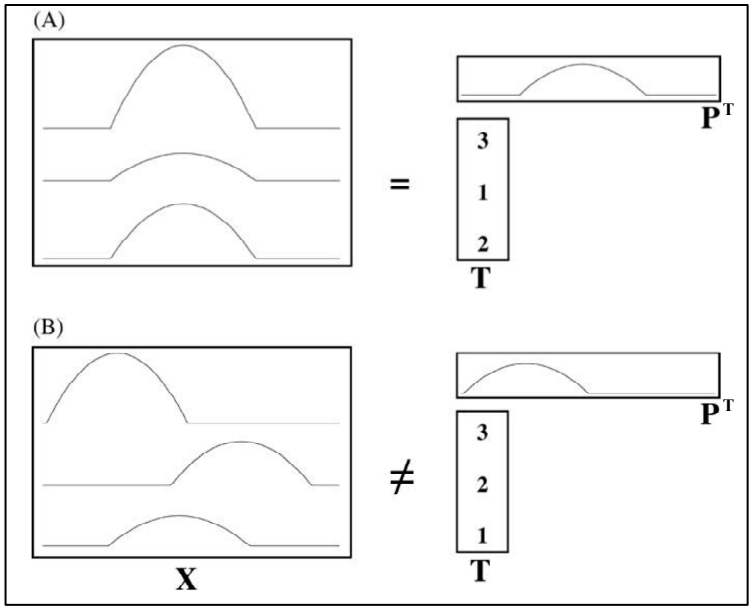


Figure 6. Illustration of the principles of data that can be modeled with a one component bilinear model (e.g. PCA or PLS) (A) (Chapter 5) and of data that show shifted profiles and cannot be modeled in a parsimonious way using a one component PCA model (B). **X** is the chromatogram holding one peak profile; **T** is the score vector and **P** the loading vector. Adapted from Skov and Bro (2005).

To solve the shifting problem; to go from B to A in Figure 6, several methods are available. For chromatographic data the Correlation Optimized Warping (COW) algorithm has proven successful as it handles non-systematic shifts efficiently [Nielsen et al., 1998; Tomasi et al., 2004]. The COW algorithm is based on aligning a sample chromatogram in the form of a vector towards a target chromatogram (i.e. a reference sample vector). This is done by piecewise linear stretching or compression in combination with interpolation, by optimizing the correlation coefficients between corresponding segments in a reference and sample chromatogram. Figure 7 illustrates the alignment of shifted chromatograms resulting in perfectly aligned chromatographic profiles. Notice the non-systematic shift as some peaks are shifted to the left and some to the right.

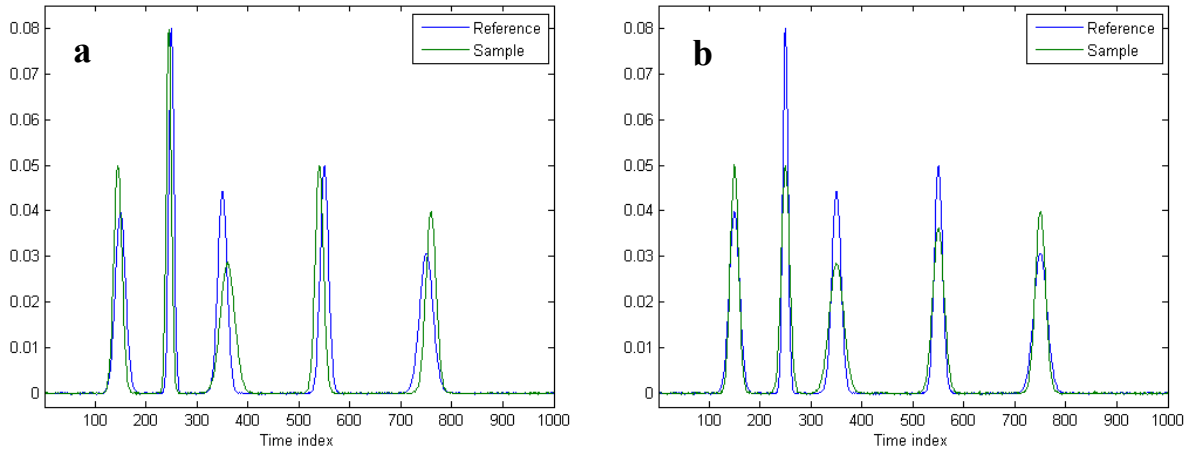


Figure 7. Alignment of chromatogram fingerprints/profiles in a data set consisting of two samples. One sample is selected as the reference chromatogram. (a) Unaligned chromatograms, (b) aligned chromatograms using e.g. Correlation Optimized Warping [Nielsen et al., 1998; Tomasi et al., 2004].

In Chapter 4, COW as well as other alignment methods will be presented and discussed in further detail with the focus on choosing proper alignment parameters and ways to evaluate the quality of the aligned data.

2.4.3 Baseline correction

Baseline correction has been an issue in chromatography for decades. Several baseline methods are available in the literature and one of the first descriptions of the importance of removing and how to remove baseline drifts was presented in 1965 [Wilson and McInnes, 1965]. They used a simple approach of integrating the area below the peak profile and then subtracting this from the overall area (peak plus baseline). Nowadays most methods are based on subtracting a fitted polynomial following the baseline curvature or by separating baseline and signal in a low-rank factor model.

So far only few baseline correction methods are implemented in commercial chromatographic software and often they tend to be too simple and generic. For chromatographic data analysis one of the most cited baseline correction tools is the asymmetric least squares smoothing by Eilers (2004). This method works by fitting an initial polynomial of a certain order to all data points in the chromatogram. By iteratively penalizing (or weighting) positive deviations (signal above the fitted polynomial) more than negative deviations (lower intensity signal plus baseline points) the polynomial will at some point approximate the baseline within a predefined limit and the resulting polynomial can be subtracted from the overall signal. This method has been shown to be efficient for chromatographic data where the baseline often can be estimated as a smooth lower-order

polynomial. The principles in penalizing (or weighting) deviations from the fitted polynomial have been utilized in many other baseline correction methods based on slightly different objective functions than presented by Eilers (2004) (Figure 8).

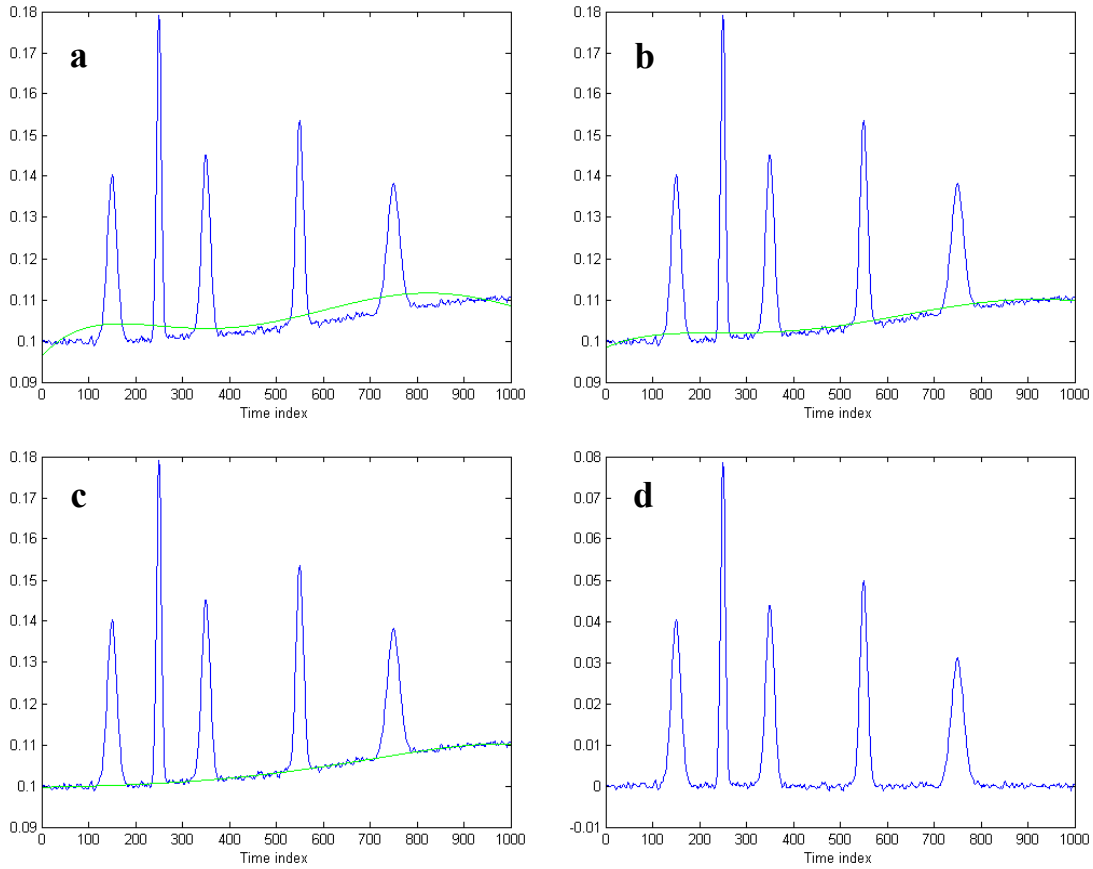


Figure 8. Principle in iterative baseline estimation using lower/higher order polynomial fit to a reduced number of data points (e.g. Eilers, 2004) in successive steps. (a-c) consecutive steps in baseline estimation pursuing that data point with signals will not influence the polynomial fit (d) baseline corrected chromatogram.

2.5 Peak co-elution/overlapping peaks

There is no need to say that the resolution of overlapping peaks is the milestone of the chromatographic analysis. As Maeder and coworkers pointed out, there are two main ways of tackling the problem; ways related to the chromatographic instrument (hardware) and to the mathematical algorithm (software) approach [Maeder and Zilian, 1988]. The *hardware* approach turns its attention to the chromatographic parameters (column characteristics, temperature gradient, mobile phase, flow etc.) to achieve perfect resolution as can be seen for the outer peaks in Figure 9. Quantitative determination of the individual analytes can simply be done by peak integration. However, this method is only valid as long as all peaks are well resolved. When peaks overlap (middle peaks in Figure 9) the *software* approach can

be used to mathematically resolve the overlapping peaks into pure peak profiles. Often the terminology changes from peak integration to peak deconvolution when mathematical peak fitting tools are employed.

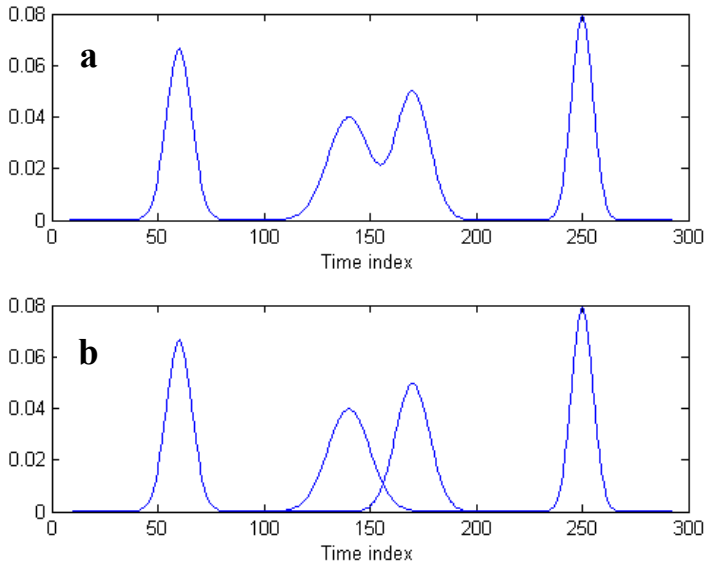


Figure 9. Illustration of overlapping/co-eluting peaks, where peak 2 and 3 overlap/co-elute (a) The overall signal (noise and baseline left out for simplification) (b) The unique parts of the co-eluted peak.

The integration of the peaks in Figure 9 is an easy task for peak 1 and 4 whereas the situation is less straightforward for peak 2 and 3. If each sample is represented by a vectorized chromatogram (one dimensional sample) the whole data matrix will be a two-way array and no additional information about the contribution from analyte 2 and 3 to the overall signal is available. However, if each sample is characterized by an additional multivariate dimension (e.g. fluorescence, UV or MS spectrum) then the additional information about each analyte can be utilized to deconvolute/resolve the overlapping peak into the unique contributions (Chapter 5).

To handle the two-way data (samples × elution times) a peak fitting approach can be used where assumptions about the peak shapes can be used to estimate the overlapping peaks. For gas chromatographic peaks, a Gaussian shape is often assumed and by changing the parameters for the involved peak functions, the overall signal can be approximated (least squares fit) in an iterative way. Stated differently, peak deconvolution consists of fitting an experimental chromatogram (or a set of them) to a linear combination of individual chromatographic peaks. Hence, a mathematical peak model is needed to describe each elementary peak. Several peak fitting algorithms and procedures are available, but they are outside the scope of this thesis. Simpler peak dividing methods are also available, but they have been shown to be extremely error prone in situations with different peak heights of the overlapping peaks (Figure 10) [Bicking, 2006a; Bicking, 2006b].

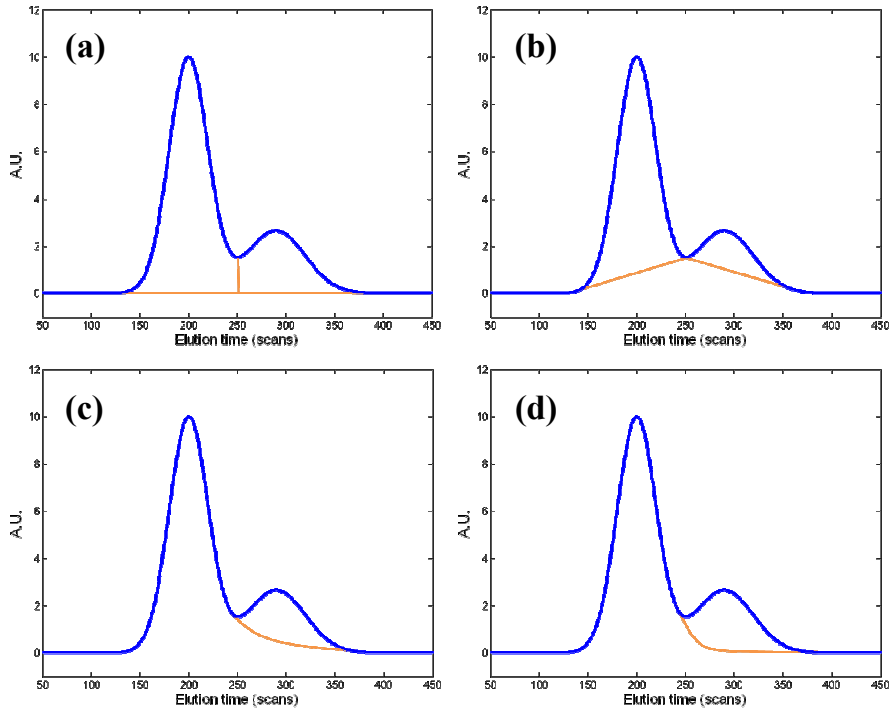


Figure 10. Common peak dividing methods applied prior to peak integration. (a) drop, (b) valley, (c) exponential skim, and (d) Gaussian skim (modified from Bicking, 2006a).

As can be seen all four approaches lead to an incorrect division of the overlapping peaks and as such wrong peak integration. This has been shown to depend primarily on the peak height ratios of the two peaks and on the asymmetry (e.g. tailing) of one or both peaks [Bicking, 2006a; Bicking, 2006b; Foley, 1987]. As will be demonstrated later (Chapter 5) the use of these rather simple peak dividing approaches (often implemented in commercial chromatographic software) may result in underestimation or overestimation of peak areas.

The methods described so far depend heavily on the peak shape when fitting either several peak functions or simply integrating using e.g. sum of data points. The peak shape will determine whether functions can be fitted appropriately and whether a peak can be split accurately into the individual contributions. Linear models like PCA and others (Chapter 5) could be used but these methods depend on peak shapes being similar across samples. Having multiple samples measured with hyphenated analytical instruments provides data of increased complexity but also data holding information which can be used to separate the mixed signal into individual contributions. For a co-eluted peak in a one sample system measured with GC-MS, Multivariate Curve Resolution (MCR) can be used to find analyte specific elution time and mass spectral profiles [de Juan and Tauler, 2007; Tauler, 1995]. Even though MCR is also a bilinear model, imposing several constraints related to the properties of analytical chemical signals (i.e. non-negativity and unimodality) true chemical features can sometimes be extracted. Having more than one sample, multiway methods have

been shown to be able to extract unique elution time and mass spectral profiles from low-rank data systems; i.e. a system of co-eluting peaks (Chapter 5). Among these methods, PARAFAC, an extension of PCA to multiway data, has proven to be suitable for aligned GC-MS and GC×GC-MS data [Bro, 1997; Bro, 2006]. If peaks are not aligned, PARAFAC2, handles the shifted behavior of similar peaks across samples by not restricting all elution time profiles to be similar in all samples, but rather that each sample can hold a unique elution time profile (i.e. different elution times and peak shapes) [Bro et al., 1999; Kiers et al., 1999]. The latter provides a real chromatographic description of data without any pre-processing steps needed. A more detailed description of these methods is given in Chapter 5.

3 Chromatographic data

IUPAC Recommendations 1993 – [Ettre, 1993]

CHROMATOGRAPHY is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (the stationary phase) while the other moves in a definite direction (the mobile phase). A graphical or other presentation of detector response, concentration of analyte in the effluent or other quantity used as a measure of effluent concentration versus effluent volume or time is called the CHROMATOGRAM.

The aim of this chapter is not to deal with how to optimize the chromatographic and mass spectral conditions for proper resolution and detection. Neither is it the idea to present chromatographic theory and instrumental setup. This can be found in more dedicated textbooks [Grob and Barry, 2006; Gross, 2006]. The focus will be on the data structure and dimensionality that provide higher order data and offer exciting possibilities with respect to peak deconvolution.

The three types of chromatographic data that will be presented shortly are:

1) GC (univariate detection)
2) GC-MS (multivariate detection)
3) GC×GC-MS (multidimensional chromatography and multivariate detection)

These techniques have many commonalities and GC (univariate detection) can simply be explained as an extracted monochannel from a GC-MS experiment (e.g. similar to Single Ion Monitoring – SIM) and as such the focus here will be on GC-MS and GC×GC-MS. The main part of this thesis will focus on GC-MS data and the data structures derived from this. The GC×GC-MS has mainly been treated in PAPER V and will be touched upon in Chapter 5.

3.1 GC-MS

Coupling or hyphenating a gas chromatographic column and a mass detector provides an analytical instrument with increased selectivity and sensitivity compared to using one of the two instruments alone. In GC-MS the hyphen indicates that the instruments are linear with respect to the coupling between the two instruments. This is the generic symbol when a chromatographic separation technique is coupled to a detector. However, as will be demonstrated in the next section the use of a × between instruments can be applied when

additional separation is achieved in the second instrument [Gross, 2006; Schoenmakers et al., 2003].

A single analysis of one sample typically involves the following steps:

1) A sample (mixture of chemical compounds) is vaporized when entering the column due to the temperature program always used for GC analysis.
2) In the column, the compounds of the mixture sample are distributed between two phases; the stationary phase and the mobile phase. Those compounds held preferentially in the stationary phase are retained longer in the system than those that are distributed selectively in the mobile phase. Due to this, solutes elutes from the system as local concentrations (when detected they are referred to as peaks) in the mobile phase in the order of their increasing distribution coefficient with respect to the stationary phase. Thus, separation is achieved.
3) Eluting compounds, resolved and separated to some degree, enter the mass detector, where ionization and fragmentation take place.
4) The fragments/ions are then separated according to their mass to charge ratio (m/z) and an electron multiplier detects every ion of the selected mass that passes through the mass analyzer.
5) The digitalized signal then consists of the number of detected ions at measured mass channels from different elution times. This can be visualized as done in Figure 11 for a small elution time region.

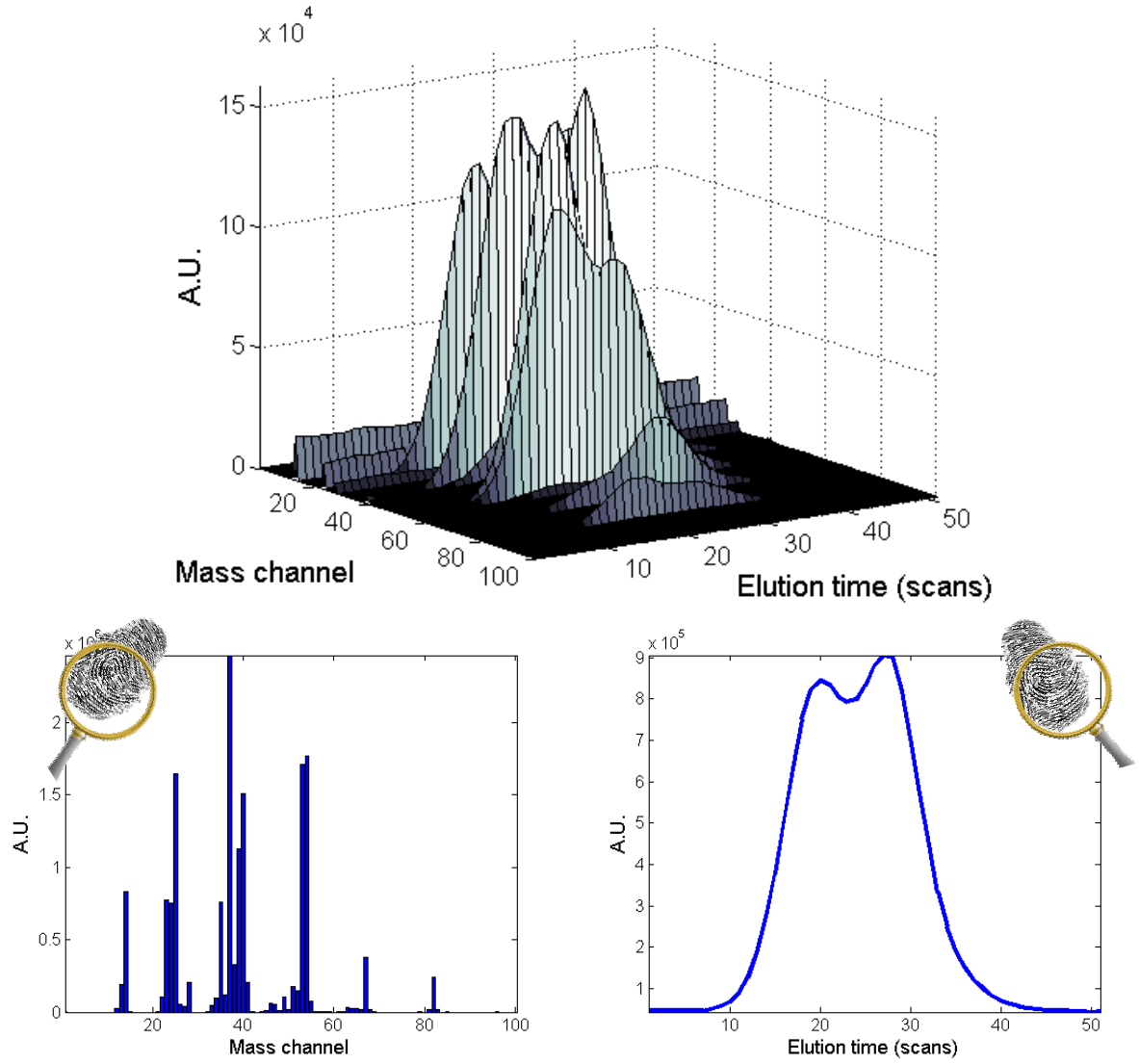


Figure 11. Illustrations of GC-MS data. TOP: *landscape* for two peaks from one sample eluting over 51 time points (scans) and detected at 100 mass channels (m/z). BOTTOM – *fingerprints*: left – mass spectral profile (summed GC dimension) and right – elution time profile (summed MS dimension, TIC chromatogram) – to be able to describe a larger pattern, the TIC chromatogram will often include all elution time scans (further discussed in Chapter 5).

As shown in Figure 11 different data structures can be extracted from a single GC-MS experiment; the intact *landscape* holding all available information, the *elution time profile* (summed MS – denoted Total Ion Count (*TIC)* chromatogram) showing that two peaks co-elute and finally, the *mass spectral profile* that for e.g. electron ionization provides a rather complex fingerprint. The terms written in italic will be used from now on to refer to these three different types of data structure.

3.2 GC×GC-MS

The analytical technique GC×GC-MS, also known as comprehensive two-dimensional gas chromatography, is a technique where two chromatographic columns are coupled through an interface system; the modulator. The separation principles taking place in each column are similar to what was explained for GC-MS, but this time the system is not linear. Rather the system consists of two columns for which orthogonal separation principles are found. In chromatography, orthogonal refers to two columns separating compounds based on different chemical properties, e.g. volatility in the first column and polarity in the second. For example, a first column coated with a common non-polar stationary phase such as methylsilicone will separate on the basis of molecular size, i.e., volatility or boiling point; whereas a second column coated with a carbowax stationary phase will separate on the basis of polarity. In general the first column is much longer than the second column as shown in Figure 12.

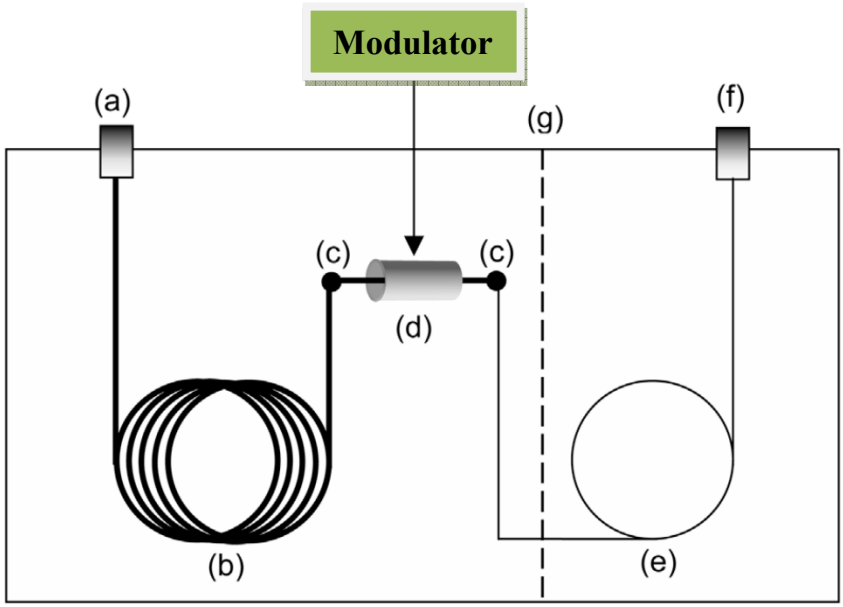


Figure 12. Block diagram of a GC×GC system. (a) injector; (b) primary column; (c) column connectors (at least one is necessary; multiple may be required depending on the exact configuration); (d) GC×GC interface; (e) secondary column; (f) detector; (g) optional division for secondary oven. Modified from Gorecki et al. (2004).

The major difference between a normal GC system and a GC×GC system is the increased peak capacity due to the additional column and the comprehensive nature of the interface system. This comprehensive nature means that all solutes eluting in the mobile phase from the first column are also introduced into the second column by means of the interface system (the modulator). Simply stated, the modulator operates by collecting a certain amount of solutes eluting in the mobile phase (effluent) before periodically introducing it to the second column. The increased peak capacity is due to the orthogonal separation principles of the

two columns and peaks not separated on the first column may be separated on the second column provided that the chemical properties of the co-eluting compounds differ with respect to the second column. Due to the very short second column a faster mass analyzer than a normal quadrupole (often used in benchtop GC-MS instruments) must be used especially when a large range of mass channels is evaluated. For most GC×GC-MS instruments a Time of Flight (TOF) mass analyzer is applied mainly due to its high data acquisition rate [Marriott and Shellie, 2002].

For detailed information on column properties, interface system and more about GC×GC-MS, the reader is referred to the literature [Adahchour et al., 2006a; Adahchour et al., 2006b; Gorecki et al., 2004; Liu and Phillips, 1991; Marriott and Shellie, 2002; Phillips and Beens, 1999; Pierce et al., 2008].

Data from GC×GC-TOFMS can be visualized in several ways, but the contour plot or color plot are highly favored as they reveal the increased peak capacity (i.e. co-eluted peaks on the first column separated on the second column) [Schoenmakers et al., 2003]. Examples of these plots are shown in Figure 13.

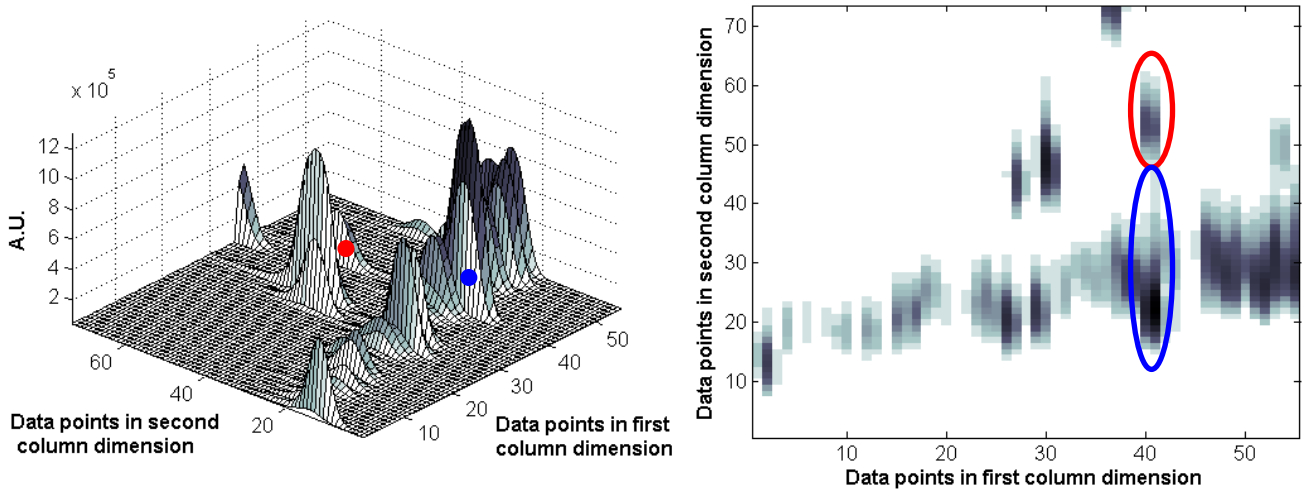


Figure 13. Visualization of data structure of GC×GC-TOFMS data where the MS dimension has been summed. LEFT: Contour plot and RIGHT: Color plot. In accordance with suggested nomenclature for GC×GC data [Schoenmakers et al., 2003].

4 Pre-processing of chromatographic data

For chromatographic data, the outcome of the pre-processing steps is data that are prepared for subsequent peak integration and/or factor modeling!

Pre-processing a chromatographic signal can be observed from two different angles: as a *chromatographer* or as a *chemometrician*. The chromatographer will use chromatographic theory to explain the effect of column characteristics, column deterioration/changes, experimental fluctuations etc., on the chromatographic signal. From this angel artifacts introduced during a chromatographic run can be evaluated and explained, and possibly used for subsequent improvement of the analytical measurements in new experiments. On the other hand, the chemometrician will often look at the chromatographic signal as a digitized signal (in the form of a vector or landscape) without considering what happened during the chromatographic run. Observed artifacts will again be considered as deviations from ideality and using mathematical methods these artifacts can be removed or handled. These two situations are of course the extremes and often knowledge from the two disciplines is combined providing a novel more powerful angel to solve a given problem.

As mentioned in Chapter 2 several methods can be applied to make the chromatographic signal ready for the subsequent data analysis. To go into details for all techniques available is not the intention in this thesis. The focus will be on selected pre-processing steps that have found their usefulness for the chromatographic discipline. The two pre-processing techniques that will be presented here are peak alignment and baseline correction (in the following peak shifts and baseline will also be denoted as artifacts). The selection of these two pre-processing methods is based on the following statements and thoughts.

1) Peak shifts and baseline offsets can be observed in all chromatograms no matter which detector is used.
2) Peak shifts and baseline can be explained in chromatographic terms and characterized and handled by mathematical terms.
3) In commercial chromatographic instruments these artifacts are handled using rather simple and not always optimal techniques.
4) The use of chemometric models requires that especially these two artifacts are dealt with to achieve rational/parsimonious solutions.
5) The efficiency in handling these artifacts can easily be visualized and/or tested with quantitative measures.
6) The mathematical methods for dealing with these problems can be automated to some extent and be made more user-friendly for chromatographers.

As mentioned before, the important information in a chromatogram lies in the areas and positions of its peaks. The area of a peak is ideally proportional to the concentration and if the peak shape is consistent between samples, the height of a peak above the baseline is also proportional to the concentration of the corresponding species. But, multivariate methods cannot always discriminate between desired variance from peak areas/height and undesired variance from baseline differences. Hence, if multivariate modeling is the next step in the data analysis then baseline correction is essential. Also, for alignment techniques a baseline contribution can hinder optimal correction as similarity measures are directly affected by the intensity of the signal (e.g. Euclidean distance). This suggests removing baseline contributions from the chromatographic signal prior to alignment and modeling. Although both steps are important the major focus will be on alignment techniques, as this has been studied extensively for chromatographic data both in the literature and during this PhD study.

After proper pre-processing, and depended on the subsequent data modeling, an intermediate step of locating the peaks can be conducted. This step serves as a further preconditioning of the data. The last part of this chapter will present some recent methods for locating peaks or peak regions that have been employed for chromatographic data.

4.1 Alignment

A data table with different information described in the same variable-entry is destructive for any linear model whether it is a classic ANOVA or a multivariate factor model. This must be corrected and we can do this with alignment techniques.

The optimal alignment technique would require only a minimal or no input by a skilled technician or scientist, be fast and would be applicable for a wide range of analytical situations without extensive customization for each of them. From the available alignment techniques some operate using the intensity (shape, profile and pattern) of the entire chromatographic *profile* in an objective way whereas others use subjectively predefined *features*. In the latter case, these can be specific (internal) standards added to the samples, specific mass channels in GC-MS data or simply peaks above a certain threshold.

For a single alignment, five different properties of the data and the alignment method are to be considered:

1) Type of data
2) Transformation of the elution time axis
3) Quality of alignment
4) Optimization of the alignment parameters
5) Choice of a reference chromatogram/landscape

Type of data

As the chromatographic separation is coupled to univariate (e.g. FID or SIM) or multivariate detectors (e.g. full scan MS, UV or DAD) the data structure changes from experiment to experiment depending on the samples analyzed. Univariate detectors and TIC chromatograms provide the simplest type of information; it is usually sufficient in GC-MS and LC-MS of low-complexity samples. For more complex samples, TIC chromatogram information is less easily interpreted because peaks can overlap, and unrelated peaks can elute at a similar elution time [Vandenbogaert et al., 2008]. If this is the case, the mass dimension can actively be used to guide the alignment procedure.

Transformation of the elution time axis

Very reproducible GC-MS data (in form of TIC chromatograms) often need only a constant or linear shift correction, e.g. stretching or shrinking of the whole elution time axis, or simply a movement of the whole chromatogram a certain integer sideways for proper alignment [van den Berg et al., 2005]. This is also known as a systematic shift. However, if the column is changed between runs, if different chromatographic columns are used or if samples are measured over a long time then a more complex shift correction is needed. This

unsystematic shift is characterized by a different degree of shift for multiple peaks across samples and can be seen as peaks shifting independent of one another in the same chromatogram.

Quality of alignment

Alignment methods typically use a sum of distances between paired features or a similarity measure to evaluate the alignment. With TIC chromatograms, the Pearson's correlation coefficient is often an attractive measure of the quality of the alignment either for single chromatograms or as an average over many chromatograms [Nielsen et al., 1998; Tomasi et al., 2004]. This measure is independent of peak intensities across samples, but will be influenced by whether the shift is observed for larger or smaller peaks. Another measure is the Euclidean distance for the similarity of two chromatographic signals [Pravdova et al., 2002b; Tomasi et al., 2004]. This measure is highly influenced by a baseline contribution or differences in peak heights.

Optimization of the alignment

The majority of alignment methods optimize the alignment based on the quality measure described above. The optimal alignment must find the best possible alignment in the shortest possible time. However, finding *the* best solution also often means much time is spent searching through all possible solutions calculated by the alignment algorithm. For this reason, several constraints can be put on the search space with respect to characteristics of the chromatographic runs. E.g. for rather homogenous data (small and rather systematic shifts) only a small part of all possible solutions may be calculated [Tomasi, 2006]. Even in the latter situation a clever/fast way of optimizing the alignment is needed. As this is an important step for making the alignment method as user-friendly as possible, this will be further discussed in this chapter.

Choice of a reference chromatogram/landscape

The choice of reference chromatogram/landscape is one of the most important aspects of the alignment methods considered here. Even with the most suitable, the most flexible and the fastest alignment algorithm – a poorly chosen reference sample can mean the difference between success and failure. Several suggestions on how to find a proper reference chromatogram depending on the type of data have been presented [Daszykowski and Walczak, 2007; PAPER I]. Among these are, the average chromatogram, the first loading of a Principal Component Analysis model, the most inter-similar chromatogram among all chromatograms or the sample run in the middle of a sequence, just to mention a few. However, the choice depends on the homogeneity of the samples, on the degree of missing peaks across chromatograms and many other things as will be addressed later.

These five characteristics can guide the selection of an appropriate alignment method, but one question that can further help the selection remains unanswered: what to do with the

aligned data? For instance, if the final analysis uses a maximum correlation/covariance criterion (e.g. PCA) then the alignment of chromatograms (fingerprints) should be performed according to this criterion. If peak wise peak integration is the final task then one might be more interested in preserving all peak area information by accepting a suboptimal alignment, but still well enough for the peak finding method to be able to match similar peaks across chromatograms.

In this chapter the focus will be on how to align GC-MS signals either as chromatographic profiles/fingerprints (GC-FID or TIC chromatogram) or as whole GC-MS landscapes (full scan). Available methods will be presented and their application areas discussed. Besides this, some attention will be paid to how to make alignment approaches more user-friendly and automated by optimization of the alignment parameters. The chapter ends with a discussion on how to select a proper reference chromatogram, as no alignment approach can do the job without careful consideration of the reference aspect [Daszykowski and Walczak, 2007; PAPER I]. For en enhanced focus on alignment, two excellent reviews describe the alignment methods in details [Tomasi, 2006; Vandenbogaert et al., 2008].

Notation

All measurement vectors will be referred to as sample chromatograms or simply *chromatograms*. The direction along which the chemical constituents elute and where alignment is required is referred to as *time* and the axis points as time index or scans. Lowercase italics are used for scalars (i.e. x) and lowercase bold for row vectors (i.e. $\mathbf{x}$), e.g. a chromatographic profile. Data matrices will be denoted with bold capital letters (i.e. $\mathbf{X}$). The ij^{th} element of a $\mathbf{X}$ is denoted $x(i,j)$, where the indices run as $i = 1,\ldots,I$ and $j = 1,\ldots,J$.

Since the alignment techniques included here are all based on time axis transformations to maximize the similarity between a target/reference chromatogram and a sample chromatogram, the following specific notation will also be used. The sample chromatogram will be denoted $x_j = x(t_j)$ and the reference chromatogram $y_j = y(t_j)$, with $t_j = j$ representing the time axis in data points/scans. In the following the length of both signals are considered to be the same so index j runs from $j = 1,\ldots,J$.

4.1.1 The warping function/path

The essence in all alignment procedures involves a transformation of the time axis of the sample chromatogram to match the reference chromatogram in the best possible way. This can be formulated in a warping function $w(t_j)$. The aim is to interpolate the sample chromatogram (x_j) to the points in the warping function $w(t_j)$ to provide the aligned chromatogram $x(w(t_j))$, which maximizes the similarity (or minimizes a distance) between the sample and the reference chromatogram (y_j) (Equation 1).

$$y_j \cong x(w(t_j)) \qquad j = 1,\ldots,J \tag{1}$$

Stated differently, we seek a relationship that associates the scan number in the sample with a scan number in the reference – this relationship can be a set of indexes or as shown an explicit function. As will be discussed later, alignment methods for chromatographic data are based on one of these two relationships. In case of a simple shift a linear or a quadratic function (Equation 2) as the applied warping function will sometimes be enough to correct for the shifted behavior.

$$j_{\text{reference}} = a_2 j_{\text{sample}}^2 + a_1 j_{\text{sample}} + a_0 \qquad j = 1,\ldots,J \tag{2}$$

where j refers to the time axis point of sample and reference chromatogram.

When a larger flexibility is needed to correct for the shifts a segmented or piecewise (local) approach can be applied. If the shift can be described by the shape of a parabola (e.g. as estimated in Equation 2) then a more flexible approach would simply be a piecewise linear approximation of this parabola. Smaller pieces/segments and high flexibility in the slope of these local linear approximations would fit better to the parabola and provide an optimal alignment. This scenario can be extended to more complex shifts, where a quadratic function is not flexible enough, but where the piecewise method can provide different flexibilities to different parts of the chromatogram to align.

Equation 1 and 2 illustrate the principles behind *parametric* alignment techniques (parameters must be optimized) and they are in contrast to the *non-parametric* methods, where a warping path is found instead of a function. Dividing the chromatograms into segments and either moving segments and/or boundaries sideways followed by an interpolation step means that segments in reference and sample chromatograms can be compared (local comparison). The warping path can then described as e.g. the new positions of the boundaries on the elution time axis providing optimal similarity between local or global structure/patterns in reference and sample chromatogram.

4.1.2 Alignment techniques used for fingerprint chromatographic data

Several papers have dealt with the comparison of dedicated alignment algorithms for chromatographic fingerprints. Among these are Correlation Optimized Warping (COW) and Dynamic Time Warping (DTW) [Tomasi et al., 2004], COW, Parametric Time Warping (PTW) and Semi-parametric Time Warping (STW) [van Nederkassel et al., 2006a], COW, PTW, DTW [Hendriks et al., 2005], COW and DTW [Pravdova et al., 2002b], COW, STW and Target Peak Alignment (TPA) [van Nederkassel et al., 2006b] and COW, DTW and PTW [Szymanska et al., 2007] – Table 2. Alignment algorithms differ mainly in four aspects; 1) the way the warping function/path is defined (non-parametrically or parametrically), 2) whether landmarks are used to guide the alignment, 3) the similarity measure that must be optimized (e.g. Euclidean distance, correlations coefficient, sum of squares of the difference between sample and reference chromatogram) and 4) the algorithmic technique that is used to find the optimal warping function or path. The reader is referred to Tomasi (2006) for a detailed description of these aspects which will not be covered in detail in this thesis.

In the remaining part of this section COW will be presented in more details, as this method has been most intensively studied both in this PhD study, but also in the literature when describing potential alignment methods for gas chromatographic data. Among the other methods some operate by principles similar to COW; the overall alignment problem is boiled down to solve local alignment problems (piecewise) in such a way that the global alignment is optimized, and they will also be given some attention.

Table 2. Presentation of selected alignment methods dedicated for profile chromatographic data.

Alignment method		Quality measure	Input parameters to consider			
Correlation Optimized Warping	COW	Correlation coefficient	Reference	Segment length	Slack size[1]	
Forshed – Piecewise alignment	F-PWA	Correlation coefficient	Reference	Segment length	Range of linearly interpolating and shifting[1]	
Pierce – Piecewise alignment	P-PWA	Correlation coefficient	Reference	Segment length	Sideways shifting[1]	
Parametric Time Warping	PTW	Sum of squares of residuals	Reference	Warping function coefficients		
Semi-parametric Time Warping	STW	Sum of squares of residuals	Reference	Warping coefficients for B-splines	Number of B-splines	Penalty term (reduce the flexibility of the warping function)

[1]This parameter describes what is done with the segments or segment boundaries in the sample chromatogram and for all methods a larger value means higher flexibility. In the following this parameters will be denoted either as flexibility or flexibility parameter.

Other alignment methods than presented in Table 2 have also been studied for chromatographic data. These include fuzzy warping where mutually corresponding peaks between reference and sample chromatogram are found and are matched and aligned using linear interpolation. The fuzzy warping has been shown to be effective especially for huge data signals like NMR spectra [Wu et al., 2006], but it has also been demonstrated for chromatographic data [Walczak and Wu, 2005]. Dynamic Time Warping was originally developed in speech recognition to align vectorized signals [Itakura, 1975; Sakoe and Chiba, 1978]. DTW aims at aligning two chromatograms by warping the time axis iteratively until an optimal match (according to a suitable measure) between the two chromatograms is found. Because of its large flexibility, DTW is widely used in many scientific fields. Several papers have described the use of DTW for chromatographic data and the overall conclusion is that DTW performs well provided that restrictions to the large flexibility are implemented [Pravdova et al., 2002b; Prince and Marcotte, 2006; Tomasi et al., 2004; Tomasi, 2006].

Correlation Optimized Warping – COW

For chromatographic profile data, COW has been demonstrated in several publications to be an efficient alignment technique [Christensen et al., 2005b; Christensen and Tomasi, 2007; Daszykowski and Walczak, 2006; Malmquist et al., 2007; Nielsen et al., 1998; PAPER I; Pravdova et al., 2002b; Szymanska et al., 2007; Tomasi et al., 2004; van Nederkassel et al., 2005; van Nederkassel et al., 2006a; van Nederkassel et al., 2006b].

Principle of COW

The Correlation Optimized Warping technique (COW) was originally introduced by Nielsen et al. (1998) as a method to correct for shifts in vectorized data signals. It 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. Without going into details, dynamic programming is a mathematical tool that works by solving combinatorial optimization problems such as finding the optimal warping path This is done by examining all the possible combinations (piecewise) of feasible transformations of the elution time axis. Details about dynamic programming can be found elsewhere [Tomasi et al., 2004; Tomasi, 2006].

A schematic illustration of COW is shown in Figure 14.

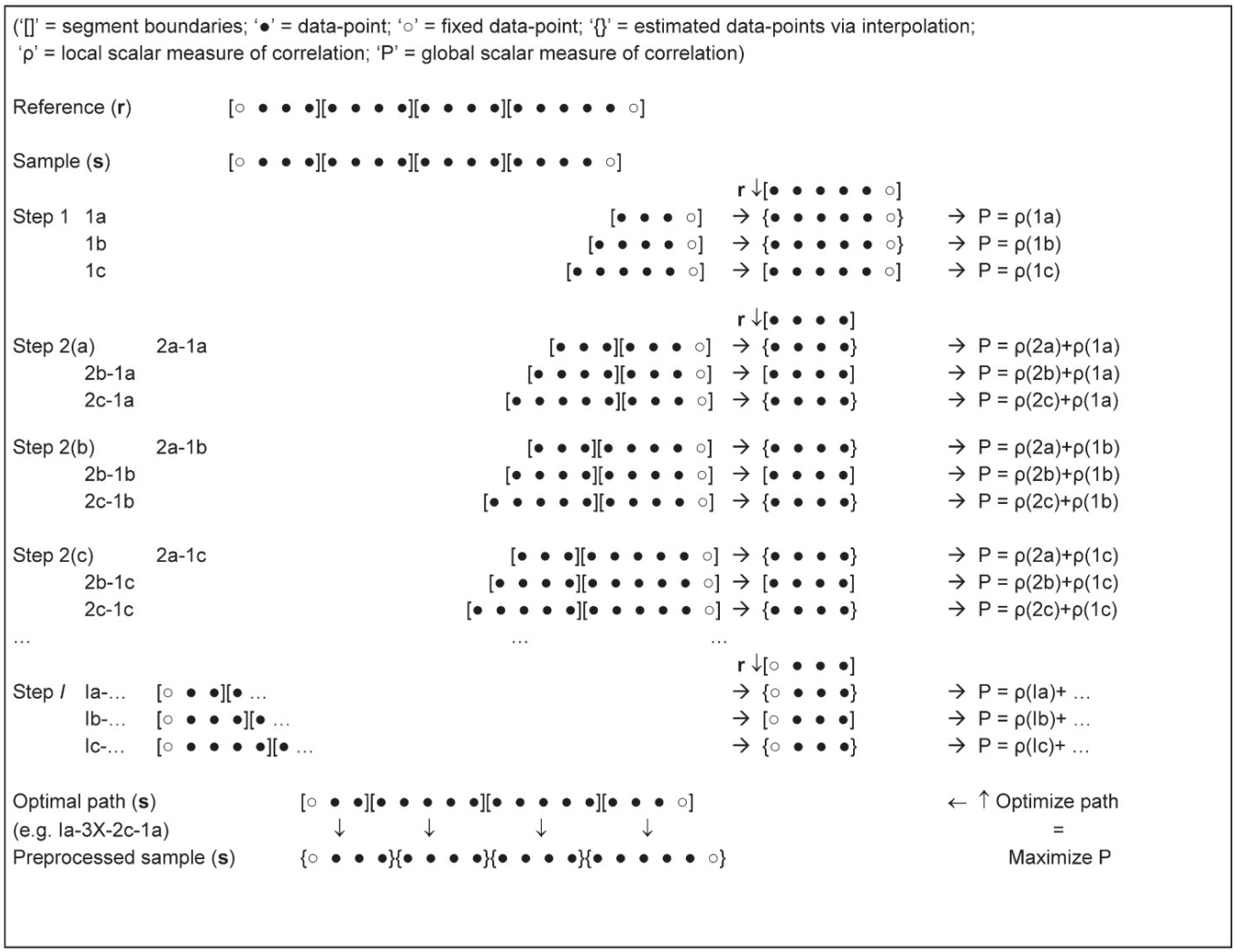


Figure 14. Schematic illustration of Correlation Optimized Warping (COW) as presented by Tomasi et al. (2004). In Figure 15 the principles of linearly interpolating the reference segment after moving a segment boundary is further visualized. Notice that in this example the segment boundary is positioned between data points, but the same explanation of COW can be used regardless of the algorithm [Nielsen et al., 1998; PAPER I; Tomasi et al., 2004; Tomasi, 2006].

The steps involved in the COW are presented below (in details according to Figure 14). The idea of the comprehensive description is to make laymen familiar with the many steps and demonstrate the rather straightforward operations in COW.

1) The reference **r** (length L_r) and sample **s** (length L_s) chromatograms are divided into a user-defined number of segments N – in the example a segment length of four is selected. The two chromatograms can have different lengths.
2) The COW algorithm starts at the end of the chromatogram and works towards the beginning. In the example a slack size (flexibility) of one is selected.
3) The outer boundaries in the outmost segments are kept at fixed positions. The left boundary of the first segment to align in the sample chromatogram is then moved one data point to the left, not moved and moved one point to the right.

4) As seen the first segment of the reference sample is one data point longer than the original segment of the sample. This is no problem, but this length determines the length of the interpolated sample segments. In this example each sample segment in **step 1** is interpolated to a length of six data points (a simple illustration of interpolation is shown in Figure 15).
5) The correlation coefficient between the three sample segments (1a, 1b, and 1c) and the reference segment is calculated and stored. In the example a segment shown as [●●●○] or [●●●●] in square brackets means no interpolation is required, whereas a segment {●●●○} or {●●●●} shown in curly brackets implies that the segment has been interpolated to fewer/more data points.
6) **Step 2** includes the second segment of the sample chromatogram. This segment is originally four data points and thus, *three* new segments of length three, four and five data points, respectively, are created. The best position of the second boundary is achieved straightforwardly by calculating the *three* correlation coefficients (**step 2**) between the second segment (interpolated to four data points) and the segment of the reference chromatogram.
7) The performance of the total warping so far is then the sum of the two correlation coefficients available in each of the now *nine* **step 2** combinations.
8) This is continued until all boundaries have been moved (notice that the last boundary is the second from the start, as the start and finish positions are fixed in COW – the empty circles in the example) and a summed correlation coefficient for all combinations has been calculated and stored.
9) From this, the *optimal warping path* can be found as the combination holding the maximum sum of correlation coefficients – stated differently to score a warping solution, an objective function, P, is constructed as the cumulative sum of the correlation coefficient of the previous sections. This solution holds the answers to what should be done with the segment boundaries (the manipulation of the time axis) and because the length of the reference segment is known these segments can be interpolated to match that length ($L_r = L_{s,aligned}$).
10) Having done this gives the best possible aligned sample chromatogram using the specific segment length and slack size (and reference chromatogram). This warping path can be utilized for other chromatograms e.g. on individual mass channels when the optimal warping path is found from the e.g. TIC chromatogram – e.g. as done by Johnson et al. (2004).
11) Some important observations can be made. The flexibility increases towards the middle of the chromatogram – this means that padding data points (e.g. white noise) in beginning and end will be needed if peaks are eluting early or late (to avoid using too high a flexibility). White noise is random noise and not correlated and as such these data points will have a very small influence on the alignment, but will increase the time for the alignment process [Fellinger A., 1998a]. For small segments a high

flexibility is simple not feasible, as candidate solutions might include overlap between successive boundaries thereby conflicting with the principles of COW.

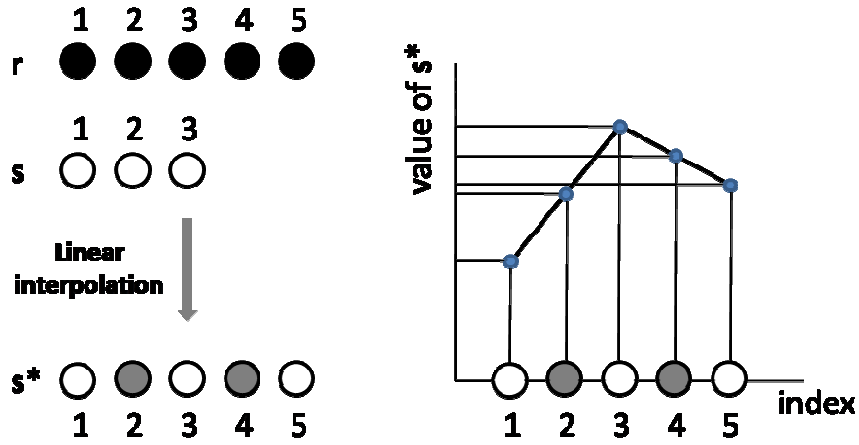


Figure 15. Principles of linear interpolation: the black circles are points of a reference signal **r**, and the white circles are the points of sample signal **s** to be aligned. To stretch the section of signal **s** to the same length as that of the corresponding section in signal **r**, the linear interpolation is performed between points 1-2 and 2-3 of signal **s** and two new points (gray circles) in positions 2 and 4 in the interpolated signal **s*** are inserted. These two new points are average values of points 1-2 and 2-3 of signal **s**, respectively [Wu et al., 2006]. As seen the shape of the values of **s** (e.g. a chromatographic peak) is maintained after interpolation – this feature is an important prerequisite for proper alignment preventing peak distortion.

Since other alignment methods used for chromatographic data have many commonalities with COW they are only shortly described in the following.

Piecewise alignment

Piecewise alignment (PWA) is a similar technique to COW in that the global alignment problem is broken down into smaller alignment problems. These local alignment problems can be solved independently (Forshed et al., 2003; Lee and Woodruff, 2004; Pierce et al., 2005; Pierce et al., 2007; Yao et al., 2007) and then put together to reconstruct the aligned chromatogram. Moreover, these methods also use the correlation coefficient between sample and reference segments as the measure of the quality of the alignment. The difference can be found in the way the global solution is found. As mentioned in the previous section, COW solves the alignment problem globally and optimal boundary positions in the first segment might not be the optimal position taking the best position for the other boundaries into consideration. This means that taking only half the chromatogram for all samples will align this part differently compared to aligning the whole chromatogram for all samples. For the PWA techniques this is not the case as each segment transformation is done independently and then put together afterwards.

Two PWA methods deserve some further attention as they are COW-like and have been shown to work well for chromatographic data. Pierce et al. (2005) presented a piecewise alignment (Pierce-PWA) method that uses a degree of class separation as a measure to optimize the alignment and from this the alignment parameters. As for COW the reference and sample chromatogram are divided into a user-defined number of segments all containing multiple chromatographic peaks. Where COW introduces a more "complicated/ sophisticated" stretching and shrinking of these segments, piecewise alignment assumes that the shifting within each segment is merely a scalar offset. In COW the boundaries between segments are moved left or right, whereas in piecewise alignment the whole segment is moved left or right (sideways shift correction). As each segment or window is moved the correlation to the segment in the reference chromatogram is calculated. The movement that provides the maximum correlation is used and the desired elution time correction is applied to the center point of the window. Corrections for areas between window midpoints and overlapping windows are linearly interpolated to yield an overall correction function that is applied to the sample vector. Since the degree of class separation (distance between replicate score values in PCA model) was used this method is semi-supervised. An unsupervised approach to the piecewise alignment method was suggested by Pierce et al. (2007) where the average correlation coefficient between target and aligned chromatograms was used as the measure of alignment quality. This resembles the measure suggested in PAPER I based on gathering most information in the first singular value for the COW algorithm.

For this piecewise alignment method only few interpolations steps are needed, reducing the computation time. However, less flexibility is achieved due to the independently aligned segments. Combined with a peak finding method ensuring that segment boundaries are not positioned within a peak this method seems to have great potential for rather systematic shifts often found in the rather reproducible GC chromatograms (reproducible in comparison to other separations techniques).

A piecewise alignment approach was also presented by Forshed et al. (2003) for NMR spectra and slightly modified by Yao et al. (2007) for chromatographic fingerprints. Here the reference and sample chromatogram are divided into segments and like Pierce et al. (2005) the segments or windows are shifted sideways in order to obtain optimal alignment between reference and sample chromatogram. In contrast to Pierce et al. (2005) this piecewise alignment is carried out using an interpolation step in each window to match the size of the reference segment. The optimal shift for each segment is determined from the correlation between segments in reference and interpolated segment in sample chromatogram. This is further illustrated in Figure 16.

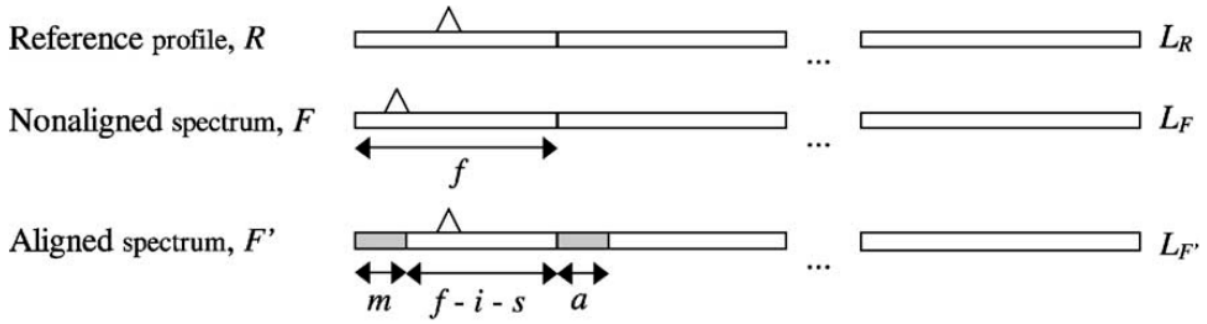


Figure 16. Principle of Forshed-PWA as presented by Forshed et al. (2003). The shifting (*s*) and interpolation (*i*) of a segment in chromatogram, *F,* is described. L_R: length of reference chromatographic profile; L_F: length of chromatographic profile to be aligned; L_F: length of the aligned chromatographic profile. By shifting and interpolation, $m = (i + s)$ points, are added to fill out the segment, and *a* points are removed from the segment to make the length equal with *f*.

Optimization of the optimal sideways movement and interpolation have been done using a Genetic Algorithm (Forshed et al., 2003) and a so-called beam search routine (Yao et al., 2007). By testing (by the author of the this thesis) this routine on chromatographic data it was found that the insertion (*m*) and removal (*a*) of data points sometimes resulted in artifacts being inserted into peak regions, but also that the method effectively and fast aligned the shift present in chromatographic data. The artifacts were observed when unfeasible input parameters (Table 2) were selected.

Parametric Time Warping – PTW

In Parametric Time Warping (PTW), a time warping function given as a second degree polynomial is calculated, where the polynomial coefficients are obtained by minimizing the sum of squares between reference and sample chromatograms (Equation 3).

$$w(t_j) = \sum_{k=0}^{K} a_k t_j^k = a_0 + a_1 t_j + a_2 t_j^2 \qquad j = 1,\ldots,J \tag{3}$$

The aim of the algorithm is to interpolate the sample chromatogram to the points in the warping function to obtain the aligned chromatogram, which is as close to the reference chromatogram as possible. One can see that if a_0, a_1 and a_2 are equal to zero, one and zero, respectively, then no alignment is performed as $w(t_j) = t_j$ for all *j*. For PTW the flexibility is restricted as only a global second order polynomial warping function can be estimated. This means that only systematic shifts (shifts in one direction) or non-complex shifts in both directions can be corrected for [van Nederkassel et al., 2006a].

Semi parametric Time Warping – STW

Semi-parametric time warping can be seen as an extension of PTW [van Nederkassel et al., 2006a; van Nederkassel et al., 2006b]. Similar to PTW, STW aims a optimizing the warping coefficients a_l in order to find a warping function that provides a minimum sum of squares. The warping function for STW is given in Equation 4.

$$w(t_j) = \sum_{l=1}^{L} a_l b_l(t_j) \qquad j = 1,\ldots,J \tag{4}$$

Here, B-splines (the lth B-spline is indicated as b_l) of varying number and extent are used instead of a polynomial of order K. The optimal coefficients are again computed by linear algebra. If the number of B-splines is too high the warping function has been demonstrated to show more variation than is justified by the data [van Nederkassel et al., 2006a]. Thus, a penalty term can be included that controls the smoothness of the fit. For STW, the number of B-splines (L) in the warping function has to be optimized and, if implemented, also the penalty term.

In a comparison study (van Nederkassel et al., 2006a) PTW has been found to be the fastest method for all data complexities and shifts investigated, but only comparable to COW and STW in alignment quality for non-complex peak shifts. For peak shifts in two directions and for real chromatograms it was found that COW and STW gave a comparable alignment as long as the number of B-splines was not too high. The advantage of STW was found to be the fast optimization of B-splines and this outperformed COW in all cases. However, COW was superior in the precision of the elution times for similar peaks after alignment. In another study by van Nederkassel et al. (2006b) similar results were found, but this time COW gave an overall better alignment, preserving the shape of the peaks better than STW. Again STW proved to be faster than COW. For the alignment with STW, the warping coefficients also need to be optimized. Similarly to PTW, this can be done iteratively and automatically [Eilers, 2004; van Nederkassel et al., 2006a]. Oppositely to PTW is that the user can fine-tune two input parameters; the number of B-splines in the warping function and the penalty term. This gives the possibility to increase the flexibility of STW. Overall parametric time warping (STW or PTW) has been only marginally used compared to COW for alignment of chromatographic data. One reason is that COW, according to this author, is easier to explain and visualize in all steps from dividing the chromatograms into segments, manipulating the elution time axis, to putting the pieces together to find the optimal warping path (Figure 14). Another reason is that COW, despite it being an often slower technique due to the many interpolation steps, has been improved significantly in speed (PAPER I) and as such can be applied to more data-rich signals (e.g. NMR spectra) as well.

In Table 3 a summary of the methods is provided, focusing on data, recommendations, concerns and more when applying the different techniques. One can see from the above

descriptions that some similarities between parametric and non-parametric methods exist. The PTW technique works globally and this is rather similar to moving the whole chromatogram sideways (two boundaries – beginning and end of chromatogram) to find optimal alignment [van den Berg et al., 2005]. Dividing the chromatograms into segments and either moving these or their boundaries sideways is a local approach that for some piecewise techniques is optimized locally (PWA) and for others globally (COW). This is comparable to STW where the coefficients of several splines must be estimated making this alignment technique focusing more on local deviations (more flexibility).

Table 3. Alignment techniques – characteristics, recommendations and concerns for different shifts for alignment methods presented (see also Table 2 and text for more details).

Alignment method		Type of shifts that can be aligned	Segments and peaks	Recommendations	Data	Concerns
Correlation Optimized Warping	COW	Complex, in two directions, unsystematic. Even though segments are set from the start non-linear shifts can still be handled if enough flexibility is provided.	Peak splitting between segments is a small problem, as boundaries are moved sideways in the alignment process.	Segment length should be larger than minimum peak width at base [Nielsen et al., 1998]. Parameters can be optimized [PAPER I]. As flexibility is increased towards the middle of the chromatogram padding white noise to the ends, can help align early and late eluting peaks.	Tested for different types of shifted data and has proven to be able to correct most shifts (see section 4.1.3 for references).	Reference sample is important. Too small segments and large flexibility can and often will change peak shapes. Time of calculation can be high due to the many interpolation steps required.
Forshed – Piecewise alignment	F-PWA	Severe and complex shifts, but within one segment only linear shifts are allowed. Segments are moved and not boundaries.	A potential problem if peaks are split by segments. But a way to control this was presented in [Forshed et al., 2003].	Since some peaks may appear in one chromatogram but not in the other, the segments must be large enough to distinguish a main pattern. The maximum range of sideways movement and linear interpolation (flexibility) for the segments must not be too large, otherwise there is a risk of fitting peaks between chromatograms which are not consistent. Shape preserving if segments are positioned correctly.	Tested for NMR and severely shifted GC data [Forshed et al., 2003; Yao et al., 2007].	Reference sample is important. Peak deformations have been observed when applying unfeasible alignment parameters.

Alignment method		Type of shifts that can be aligned	Segments and peaks	Recommendations	Data	Concerns
Pierce – Piecewise alignment	P-PWA	Severe and complex shifts, but within one segment only linear shifts are allowed. Segments are moved and not boundaries.	A potential problem if peaks are split between segments. But, the F-PWA method could be used.	Flexibility should be equal to or greater than the largest amount of elution time shifting. Shape preserving if segments are positioned correctly.	Tested for severely shifted data where peaks are shifted past neighbor peaks across chromatograms [Pierce et al., 2007].	Reference sample is important.
Parametric Time Warping	PTW	Non-complex shifts, in one direction or systematic shifts in two directions.	No segments needed, as a global warping function is applied.	Suitable for small shifts that are systematic over the entire chromatogram (quadratic function offering low flexibility). If only linear shift then an even more rigid global alignment method can be used [van den Berg et al., 2005]. Fast method.	Tested for many types of shift, but only really suitable for small systematic shifts [Szymanska et al., 2007; van Nederkassel et al., 2006a].	Reference sample is important. Low flexibility as only quadratic warping function.
Semi-parametric Time Warping	STW	Complex, in two directions, unsystematic shifts.	Number of splines must be set.	Optimizing penalty term and number of splines in the alignment is necessary to avoid peak shape changes [van Nederkassel et al., 2006a].	Tested for many types of shifts and provides good alignments in most cases [van Nederkassel et al., 2006a; van Nederkassel et al., 2006b].	Reference sample is important. Too many B-splines will result in too much flexibility and changes in peak shape. This also depends on the reference selected.

4.1.3 Optimization of the alignment parameters

In all alignment methods presented and used in the literature optimization of the alignment are done to find the best solution from a given set of input parameters (i.e. for COW finding optimal new segment boundary positions given the initial positions and flexibility using dynamic programming). Other optimization methods are beam search or genetic algorithm for the PWA approach and a gradient search for PTW and STW. For a detailed description of optimization methods the reader is referred to Tomasi (2006).

Another way of looking at optimization is to use the best aligned chromatographic profiles from a given set of input parameters and then evaluate these based on different measures applied in the alignment algorithm (optimization). The efficiency of the alignment procedure or the quality of the alignment is often evaluated mathematically using the correlation between the aligned chromatograms and visually to detect possible peak shape deformations [van Nederkassel et al., 2006a]. However, in complex chromatograms the latter is a rather time consuming job and a suggestion to solve this problem has long been advocated for.

With respect to the alignment of chromatographic profiles, two important aspects must be considered to get as optimal an alignment of the chromatographic profiles as possible.

1) The quality of the alignment – how synchronized are the chromatograms
2) The shape preserving effect – what is the change in area under the peaks or in peak shapes in the chromatograms

As shown in Table 2 the quality of the alignment approach is always determined using a certain measure (Euclidean distance, correlation coefficient etc.), which quantifies the goodness of the alignment using a certain combination of input parameters. This means that of the tested combinations of input parameters, terms or coefficients, the ones that provide the best similarity (between two or more chromatograms) will always be the proposed "parameters to use" in the given alignment routine. However, another aspect of the total alignment approach is what happens to the chromatographic signal after alignment, especially for the methods that include an interpolation step (COW, PWA) or which have an uncontrolled flexibility if the solution is not constrained (STW). Any change in the chromatographic signal, which is not along the elution time axis, is per definition a *destructive change* and is most often characterized by peak deformations or changes in peak area.

The importance of considering these potential changes in the chromatographic signal has been reported in several paper [Bylund et al., 2002; Forshed et al., 2003; Pravdova et al., 2002b; Tomasi et al., 2004; Tomasi, 2006; Torgrip et al., 2003; van Nederkassel et al., 2006a]. For all these techniques the combination of small segments and a large flexibility is

the key reason to for undesired changes in the chromatographic signal. However, another key factor is present for all the techniques mentioned so far – the reference chromatogram. Together with the other input parameters, the reference sample is an equally important parameter, which has also been reported for many alignment approaches [Daszykowski and Walczak, 2007; PAPER I; Vandenbogaert et al., 2008].

For the piecewise alignment algorithms presented in Table 2 and Table 3 (in the following focusing mainly on COW) three input parameters must be considered and set:

1) The segment length or number of segments
2) The flexibility (i.e. slack size, sideways movement)
3) The reference chromatogram

Segment length and flexibility

The quality of the alignment has been investigated in several papers using COW where some report only the combination of segment length and slack size applied found by a trial and error approach while others propose an optimization scheme [Bassompierre et al., 2007; Christensen et al., 2005b; Christensen et al., 2005a; Daszykowski and Walczak, 2007; Debeljak et al., 2005; Malmquist et al., 2007; Pravdova et al., 2002b; Schmidt et al., 2008; Tomasi et al., 2004; van Nederkassel et al., 2005].

In a study on chemical fingerprinting of petroleum biomarkers Christensen et al. (2005b) put forward the alignment parameters that were found optimal and used in the end for the data at hand. However, they did also discuss the use of an optimization method. Christensen et al. (2005b) suggested that the optimal choice of segment length and slack size was the one that maximized the first singular value. Instead of using the whole data matrix as basis for an SVD (Singular Value Decomposition – like PCA, Chapter 5) only replicated samples were used. Analytical replicates of a sample can despite changes in sensitivity along the retention time, variations in the injected sample volumes, baseline drifts, retention time shifts, alterations in peak-shape and random noise, be expected to be identical. In (van Nederkassel et al., 2006a; van Nederkassel et al., 2006b) several combinations of segment length and slack size were evaluated. They focused on finding the largest segment length and smallest slack size providing the highest possible correlation coefficient between two chromatograms. This criterion was selected to get the best alignment in the shortest possible computation time. The quality of the alignment was inspected visually plotting all chromatograms together, whereas peak shape deformation was inspected on individual aligned chromatograms. Nielsen et al. (1998) used two approaches to find optimal warping parameters. In the first approach the segment length (N) was varied while the slack size (t) was kept at a constant fraction of the value of the segment length ($N/t = 5$). In the second approach the segment length was kept constant and the flexibility varied. They concluded that the optimal segment length was approximately the peak width in data points for the

smallest peak of interest. Changing the flexibility, while keeping the segment length intact, showed that similar alignment quality was achieved. Only a small improvement in alignment was observed for higher flexibility. To avoid deformation of peaks to match features that they do not resemble originally, it was proposed to use the lowest slack size providing proper alignment.

Reference selection

The selection of the reference chromatogram gets little attention compared to the segment length and flexibility parameter (slack size, sideways movement, ranges of shift, etc.). As nearly all alignment quality measures are calculated between a reference and a sample chromatogram this is an important part of the alignment method and deserves more attention. The optimal reference sample should fulfill one or more of the following statements.

1) Be representative
2) Be complete – contain as many common peaks as possible
3) Be reproducible
4) Real – contain the same noise and background as sample chromatograms
5) Be clean – no interferences

Nevertheless, problematic issues for reference chromatograms must also be considered.

A. Samples in very different concentrations can have peaks shapes deviating from ideality (e.g. tailing peaks for overloaded columns) for some samples and others which are Gaussian shaped. In this case, extreme cautiousness is required especially for alignment methods like COW where several interpolation steps are included.
B. Using the average chromatogram as reference sample could be an interesting approach to create a reference containing all peak features. But due to the shifted peaks, the peaks in the reference chromatogram will be wider and a peak shape change for shoulder peaks and overlapping peaks will most likely be observed (Figure 17).
C. Different samples (e.g. different classes) will be impossible to align using the same reference sample. This has been addressed further by Daszykowski and Walczak (2007).

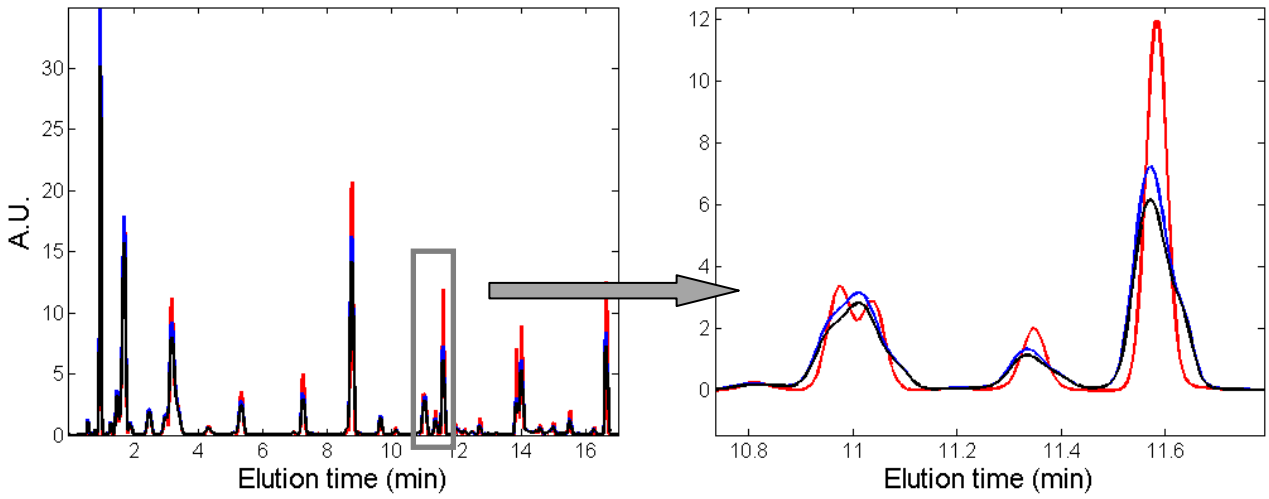


Figure 17. Three possible reference chromatograms. **BLACK**: mean chromatogram, **BLUE**: Loading of first principal component in PCA – see Chapter 5 for more on PCA, and **RED**: Real sample selected e.g. the most similar chromatogram among all.

Among popular references to use are the chromatogram in the middle of the sequence run (Bylund et al., 2002; Christensen et al., 2005b) or the most representative chromatogram containing the highest number of common chemical constituents (i.e. peaks) [Pravdova et al., 2002b; Tomasi et al., 2004; Vandenbogaert et al., 2008]. Daszykowski and Walczak (2007) came to the conclusion that the best results were obtained using the chromatogram with the highest mean correlation coefficient with respect to the remaining chromatograms as the reference (Figure 17). A similar approach using the sample providing the highest value of the product of correlation coefficients between each sample and the remaining has also been presented [PAPER I].

More dedicated optimization routines

To improve the alignment even further the optimal solution from several combinations of input parameters can be compared. Skov et al. (PAPER I) presented a simplex coordinate optimization (gradient search) routine of the two input parameters for COW – segment length and slack size. In this method it is possible to optimize the segment length and slack size for a given alignment task to remove the time consuming trial and error step. The optimization routine includes two measures: one for similarity (termed *simplicity*) and one for peak shape changes (termed *peak factor*) that are weighted equally in the overall expression that is optimized – the warping effect (Equation 5). See PAPER I for a detailed description on how to calculate these measures.

$$\textbf{Warping effect} = \lambda_1 \times \textbf{Simplicity} + \lambda_2 \times \textbf{Peak factor} \qquad (5)$$

One simplicity value and one peak factor value are obtained for one combination of segment length and slack size, and they are calculated from the best aligned chromatograms given these input parameters. The simplex coordinate optimization searches within a user-defined search space spanned by the selected segment length and slack size boundaries (Figure 18).

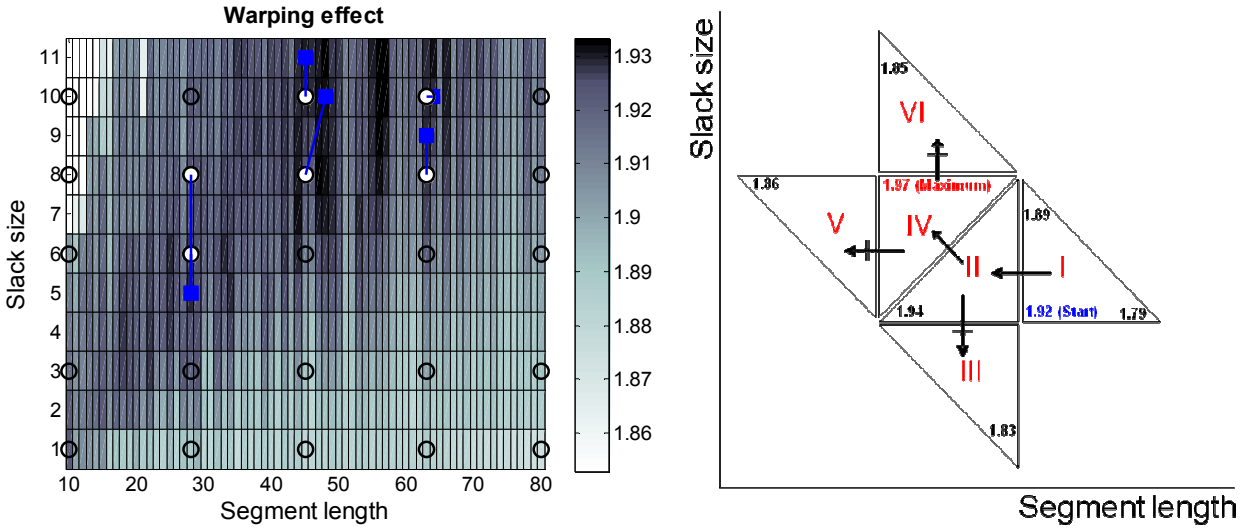


Figure 18. Example of the search space (Warping effect) for an optimization of COW input parameters with segment length from 10 to 80 and slack size from 1 to 10. Here the best alignment from twenty five combinations of segment length and slack size (open circles) are found and from these, the six largest (white circles) are selected as starting points for the simplex optimization (right plot). After optimization, one maximum Warping effect value is found from the six end points (blue squares). Notice that only the initial twenty five combinations and the combinations visited during optimization need to be calculated. Here all combinations have been calculated to illustrate that certain parts of the search space are infeasible (light colored areas). Figures can be found in PAPER I.

By using this optimization routine the number of points (combinations of segment length and slack size) visited in the search space is significantly reduced. The global optimum (maximum Warping effect value) might not be found, but it has been shown that local near-optimum combinations are satisfactory and no difference can be seen either visually or in subsequent factor models [PAPER I]. By simply setting the search space boundaries an appropriate alignment of the data at hand is achieved. The reader is referred to PAPER I for further examples and also illustrations of selecting improper alignment parameters. Several calculations must still be done compared to just one combination of alignment parameters and the acceptance of this optimization routine thus depends severely on the speed of the alignment algorithm.

In the original equation the two terms λ_1 and λ_2 are both set to one. As will be discussed later if profile similarity is more important than preserving quantitative peak information or vice versa, then user-specified weighting terms can help the optimization routine to search for a solution fulfilling this request. Another way of putting more weight on the similarity

between chromatogram was addressed in [Nielsen et al., 1999; Nielsen et al., 1998]. As the quality measure of alignment they used the cubed correlation coefficient, r^3 instead of r favoring the highest correlation values relative to the lower ones. This makes COW prefer a small number of well aligned signals over many poorer alignments. The method presented in PAPER I with evenly weighted terms (Equation 5) was tested by Christensen and Tomasi (2007) and Schmidt (2008) and good results were achieved. Schmidt (2008) used this routine for data, which had already been aligned for a different study where the optimal alignment parameters were found from a trial and error approach. It was found that an almost similar alignment was achieved, but the peak shapes were better retained using the solution suggested after optimization.

4.1.4 Further improvement of alignment

With severe shifts in data, perfect alignment can be difficult to achieve in one alignment sequence unless a large flexibility is allowed. As mentioned, this can (and often will) lead to changes in peak shape and from this an increased risk of incorrect peak area quantification. For COW, the time of calculation has been reported to increase by the square of the slack size [Nielsen et al., 1998]. In alignment studies the final combination of segment length and slack size was also chosen as the one providing best alignment in shortest possible time preserving peak profiles (i.e. shapes) [Nielsen et al., 1998; van Nederkassel et al., 2006a; van Nederkassel et al., 2006b]

As time consumption is of minor importance with the personal computers of today, the change in peak shape is a much more challenging and vital aspect. As mentioned above, the optimization of alignment parameters using a combined measure of both alignment quality and peak shape/area preservation (PAPER I) can help to avoid this problem. However, in some situations the selection of parameters cannot at the same time provide excellent alignment and intact peak area/shape in one alignment run. In this situation several alignment runs can be applied. Sadygov et al. (2006) suggested that a crude alignment was performed initially to align major features in data and to save time in the following more detailed alignment run. Their method was applied for full scan LC-MS landscapes and the initial crude alignment method (performed on TIC chromatogram) resulted in a reduction in the size of the correlation matrix needed for the remaining alignment. Likewise, Malmquist and Danielsson (1994) used the maximum covariance found from the cross-correlation between two chromatograms as a rigid initial alignment. This method was found to work well for similar samples meaning that the largest peak contributing most to the covariance measure should be found in all samples [Jonsson et al., 2004].

For other alignment methods the implementation of an initial crude alignment step has the potential to save both time and reduce the flexibility needed and thus, minimize the risk of changes in the chromatographic profile. Having said this, it is also important to state that the

effect will be most significant if some structure in the global peak shifts is observed – if peaks are shifted both left and right throughout the chromatogram the crude alignment might have no effect. The most simple initial alignment technique is when considering the whole chromatogram as one segment and then moving this segment sideways a certain number of data points while calculating the correlation to a reference chromatogram [van den Berg et al., 2005]. Although this requires the insertion of few interpolated or just replicated data points in the beginning/end of the chromatogram the increase in calculation time when changing the slack size is insignificant.

To generalize, different scenarios can be established for performing "perfect" alignment (aligned and preserved data) in the shortest possible time:

1) One alignment run with high enough flexibility to ensure perfect alignment, but preserving the chromatographic profiles
    a. Issues: time consumption, how to preserve peak shapes. Visually (van Nederkassel et al., 2006a; van Nederkassel et al., 2006b) or using quantitative measures [PAPER I].
2) Two alignment runs. If some systematic behavior in the shifts is present, the first run should be crude/simple to align main features followed by a more semi-flexible method that allows enough flexibility to local regions of the pre-aligned data.
    a. Issues: depending on the systematic shift present, the time consumption can still be an issue. Preservation of peak shapes will be less important as less flexibility is allowed for.
3) Several alignment runs can also be applied where each run is limited in flexibility (e.g. for COW a slack size of one or two). In each step an improvement in similarity between chromatograms is important as long as the peak profile is preserved. This approach will align mainly local features in the chromatograms in each run, but over many runs, global alignment will be obtained. The restricted flexibility causes each run to be rather fast, but also that several runs are needed before a perfect alignment is achieved. The latter can be controlled by some convergence criteria meaning that alignment is continued until only small differences in similarity are observed.
    a. Issues: depending on shifts observed in original chromatograms the number of runs necessary can be rather large. The changes in peak shapes will be limited from run to run, but small changes in each of several runs can be a problem.
4) Combinations of the above – e.g. initial crude alignment followed by several restricted alignment runs.

4.1.5 Alignment techniques used for GC-MS landscapes

In aligning measurements from hyphenated separation techniques (e.g. GC-MS landscapes) additional information is available from the spectral dimension (e.g. UV, MS and

fluorescence). This can be used actively or passively in the alignment procedure. **Passive** means that the spectral dimension is summed/collapsed and vector based alignment techniques, as the ones presented so far, can be used. Doing this it is assumed that no shifts are observed in the spectral dimension, which is a valid assumption for most desktop analytical low resolution instruments (e.g. standard Quadrupole mass analyzer). **Active** means that the spectral dimension is used when aligning, either as an additional tool in vector based alignment or included in the alignment quality measure (e.g. matrix correlation). A simple illustration of active/passive use of the spectral information for an alignment technique based on the correlation coefficient as the quality measure is shown in Figure 19. The example especially highlights the problems when shifts in overlapping peaks and/or complex peak systems with different peak intensities are present.

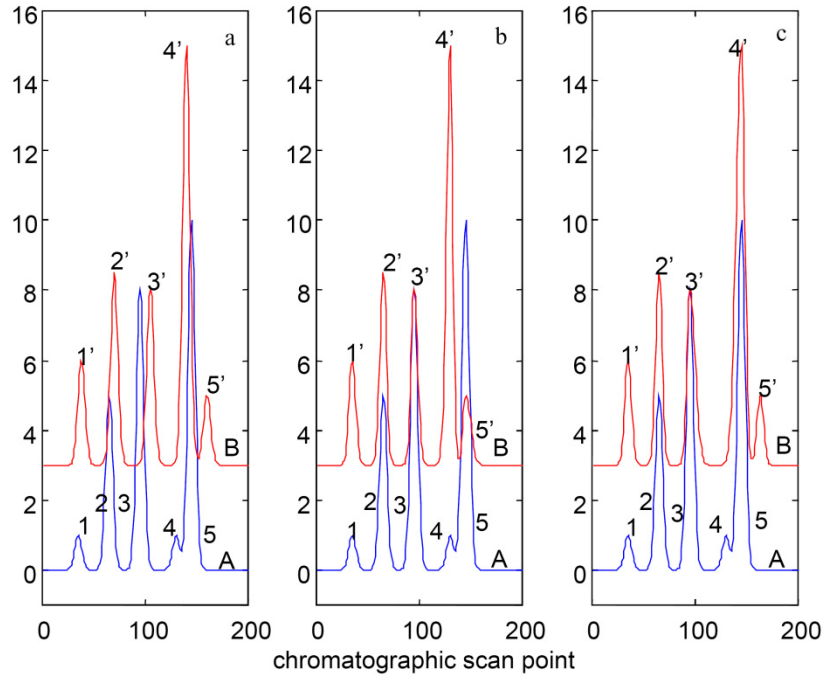


Figure 19. The use of correlation coefficient as quality measure for alignment for one dimensional (e.g. GC-FID) and two-dimensional (e.g. GC-MS) chromatographic analysis. (a) before alignment, (b) after alignment using the mass spectral dimension as guide for proper alignment - correlation coefficient of 0.57 and (c) after alignment using only the vectorized signal – correlation coefficient of 0.77. Peaks 1,2,3,4 and 5 (Reference) should match peak 1',2',3',4' and 5' (Sample). Modified according to Gong et al. (2004).

As seen, larger peaks influence the overall correlation coefficient more than smaller peaks. Nielsen et al. (1998) reported that the correlation coefficient calculated for a large segment containing both small and large unaligned peaks would measure the quality of the alignment of the larger peak compared to the smaller peaks. This is due to the intrinsic properties of the Pearson correlation coefficient used in the alignment routine. In the example above the use of just the correlation coefficient gives the aligned data in (c) where the large peaks 5 and 4' from reference and sample, respectively, are aligned incorrectly based on the chemical

knowledge (e.g. mass spectra). In (b) the mass spectral dimension is actively used to avoid aligning large peaks incorrectly (this will also affect small peaks) and now all five peaks match between reference and sample chromatogram. It seems necessary and appropriate to correct the elution time shifts by making full use of the chromatographic and spectral information provided by hyphenated chromatography.

Passive

Very often only the TIC chromatogram (section 4.1.3) is aligned and the optimal warping path/function found is then used to align individual SIC/mass channels. This is an efficient technique and treats all spectral channels in a similar fashion [Bylund et al., 2002]. However, this passive approach is severely affected by the potential problem described in Figure 19.

Active

For some alignment methods the spectral information can be used to select similar features from reference and sample chromatogram, but the alignment is still performed on the vectorized signal. This can be done by comparing e.g. the obtained mass spectrum at certain scans. By forcing similar features to match between chromatograms not only based on unsupervised correlation coefficient calculations, a supervised alignment can be carried out still operating along the elution time axis. Krebs et al. (2006) presented a method for alignment of GC–MS data that identifies landmarks (peaks in the TIC chromatogram that are above a selected threshold) in the data, comparing them across samples, and aligning only those determined to be the same by a high correlation of the peaks in the m/z dimension. As the misalignment is generally not a simple shift, a nonlinear cubic spline function is used to interpolate the elution time axis between the landmarks. This also offers the advantage of using information on each side of the matched landmarks to determine the functional fit. They define “landmarks” to be peaks above a threshold in the TIC chromatogram, and match those landmarks between two experiments if the correlation score of the corresponding mass spectra exceeds 0.99.

A more mathematical approach was described by Gong et al. (2004). Here each peak in individual sample chromatograms was resolved into pure chromatograms and spectra (see Chapter 5 for more on peak deconvolution), providing qualitative and quantitative measures of each peak in each sample. By selecting marker compounds (landmarks) known to be a specific chemical compound present in all samples they used these actively in the alignment by interpolation between these landmarks as described above. This approach requires that marker compounds are above a certain threshold in all samples. Xu et al. (2006) used a similar approach based on local factor analyses to find similar target peak(s) between chromatograms. Dividing the chromatogram into segments they aligned the most intense feature(s) in these segments. For linear shifting among chromatograms, two target components were found adequate for a good alignment. However, more target peaks are

usually necessary to better fit a non-linear shift or instead of linear interpolation a cubic spline interpolation could be used between the target peaks.

In the original COW paper (Nielsen et al., 1998) a way of including several mass channels in the correlation coefficient expression was presented. For one GC-MS landscape each mass channel was treated as a single observation (row) measured at several elution times. By taking the average of the row-wise (mass channel wise) calculated correlation coefficients between the reference and a sample the mass spectral information was actively included in the quality measure of alignment. Doing this, more stable alignments were achieved with respect to variation of free parameters such as the maximum allowable shift correction. The same increased stability in the alignment routine was reported in Szymanska et al. (2007), where several wavelengths were considered/included in the alignment of Capillary Electrophoresis (CE-DAD) profiles. A similar method which hasn't been published works by calculating the matrix correlation coefficient, which should also stabilize the alignment and lower the chance of aligning peaks with different mass spectra. A recent study for alignment of LC-MS data provided a quality measure not based on the correlation coefficient, but on the overlap integral for peaks pairs (peaks to be aligned) fitted by two 2-D Gaussian peak functions. This measure serves as the figure of merit for 1-D warping and feeds directly into the COW algorithm. This approach works in local regions in the LC-MS landscapes (but could be applied for GC-MS data as well) and a peak finding algorithm is thus an essential initial step [Suits et al., 2008].

4.2 Baseline correction

In all signals measured some degree of non-analytical contribution will be present, the baseline offset. As the ratio signal$_{analyte}$/signal$_{baseline}$ changes for different concentrations any calibration model will be affected. Thus, baseline correction is needed.

In Chapter 2 the principles of removing a baseline by fitting a polynomial through a certain number of data points was visualized. Here the main baseline correction methods proven to be efficient for chromatographic data will be touched upon. Baseline correction in chromatography is commonly employed to eliminate interferences due to baseline drift, column bleed and overlap of broad, poorly defined peaks attributable to complicated sample matrices. The most common baseline issue for GC analysis is when using a temperature program during analyses to handle the general elution problem. The temperature is raised during and/or at the end of run to make sure less volatile compounds are vaporized, but this result in an increasing baseline [Hinshaw, 2001]. A drifting base line during the elution of a GC single peak can prevent accurate quantitative analysis. A bleeding column, introduces additional mass spectral signals that complicate or even preclude correct identification of the eluting component and prevent accurate quantification in the chromatographic domain.

In chromatography the contributions to the measured signal can be expressed in the following equation (illustrated in Figure 4 in Chapter 2):

$$y(x_j) = b(x_j) + s(x_j) + \varepsilon_j \qquad j = 1,\ldots,J \tag{6}$$

where, $y(x_j)$ is the measured signal, $s(x_j)$ is the analytical relevant signal, $b(x_j)$ is the baseline, and ε_j is the noise/measurement error at elution time j. From this it can be seen that being able to find the baseline term and then subtracting this would result in a baseline corrected signal.

Typically, the univariate response from a single channel sensor is corrected by baseline skimming or by fitting polynomial functions to the baseline in the vicinity of the peak to be corrected. Or the polynomial can be fitted through a predefined number of base points either with a global polynomial or local polynomial fit. In general, global polynomial fit requires higher order polynomials whereas local fit requires lower order polynomials. However, local polynomials require that intersections are continuous, which is not accomplished using ordinary polynomials with no restrictions. The latter can be solved using splines (e.g. cubic splines) as these are piecewise linear functions which have continuous derivatives up to the order of the spline [Fellinger A., 1998c].

As is the case for alignment, baseline correction methods for both vectorized data (i.e. fingerprints/profiles) and data for hyphenated instruments (e.g. GC-MS landscapes) are available. For alignment, no shift is assumed in the mass spectral direction and the found alignment parameters (warping path) from TIC chromatogram alignment across samples can be applied for individual mass channels. This is not valid for baseline correction, as individual mass channels will have different baseline contributions due to the column material only giving rise to fragments in certain mass channels. On the other hand this can also be convenient, as knowledge about the column material can guide the selection of which mass channels to baseline correct or maybe remove/ignore [Dixon et al., 2006; Willse et al., 2005].

4.2.1 Methods for fingerprint chromatographic data

One way of removing baseline from a vectorized signal was shown in Chapter 2 (Figure 8). This approach was based on fitting a global polynomial and, through an iterative routine, down-weighting points belonging to the signal. Then the baseline was constructed and subtracted from the original signal. Two similar approaches will be shortly presented here. The first was presented by Gan et al. (2006). They also fitted a polynomial of a user-defined order to the vectorized signal and by setting signal intensities above the fitted polynomial

equal to the polynomial they iteratively found the final baseline. The idea has been illustrated in Figure 20.

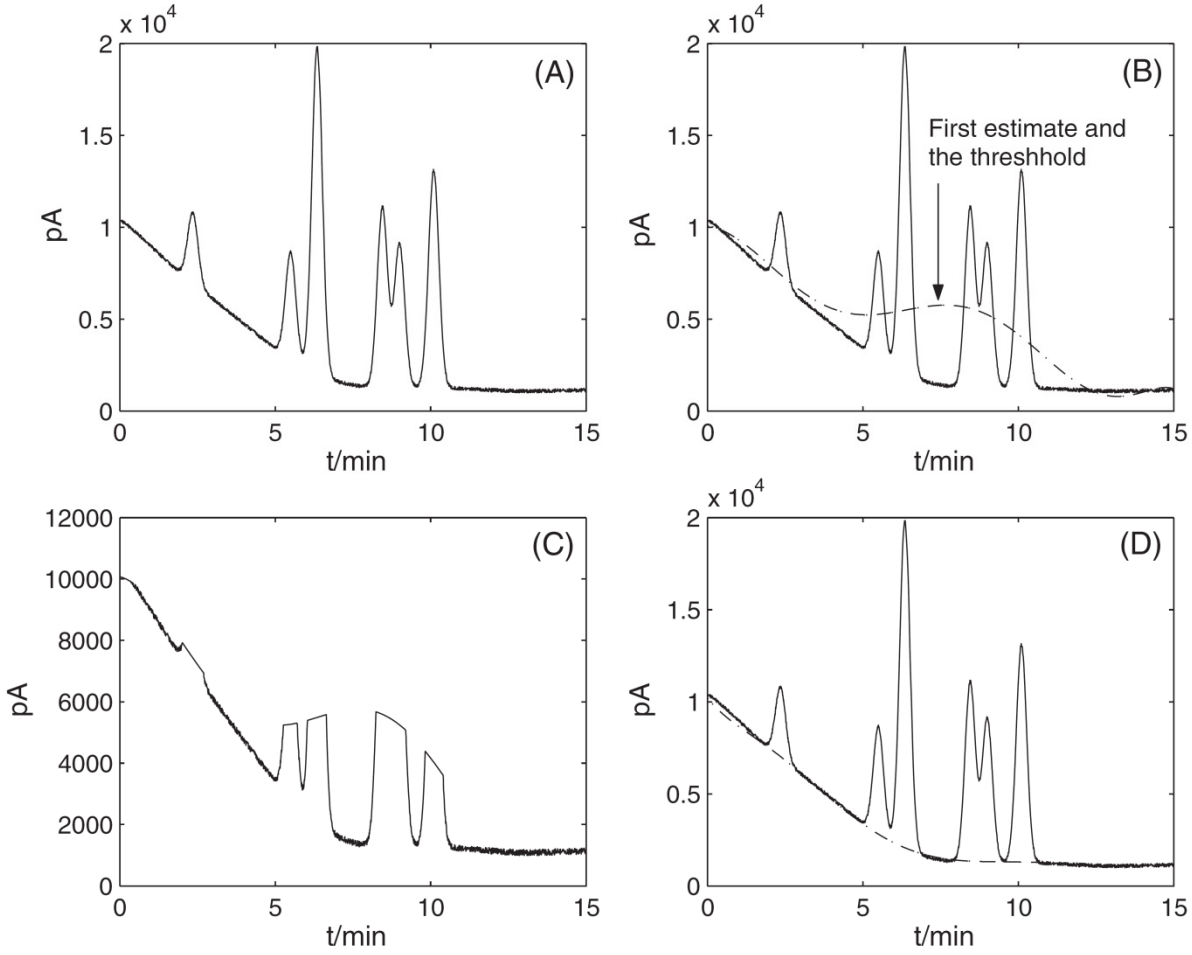


Figure 20. Illustration of baseline correction presented by Gan et al., 2006. (A) Analytical Signal; (B) First polynomial fitting; (C) Signals above polynomial set equal to the polynomial and another polynomial is fitted and (D); The final estimated baseline. Here an order of seven was used for the polynomial. Modified from Gan et al. (2006).

This method showed good potential for baseline separated peaks, and also for less severe co-eluted peaks systems. With many successive peaks overlapping, meaning no baseline points in between peaks, this routine still performed well [Gan et al., 2006].

Another method (due to van den Berg (2008) – unpublished work) operates in local regions of the chromatogram and uses B-splines that are constructed from polynomial pieces joined at certain values of x (x_j), the knots. Thus, instead of fitting local polynomials, which are not necessarily continuous between regions, splines offer a smooth baseline. The theory of splines will not be discussed here but B-splines offer a high flexibility in the fitted signal but too many knots can lead to overfitting. A simple illustration of this method is presented in Figure 21.

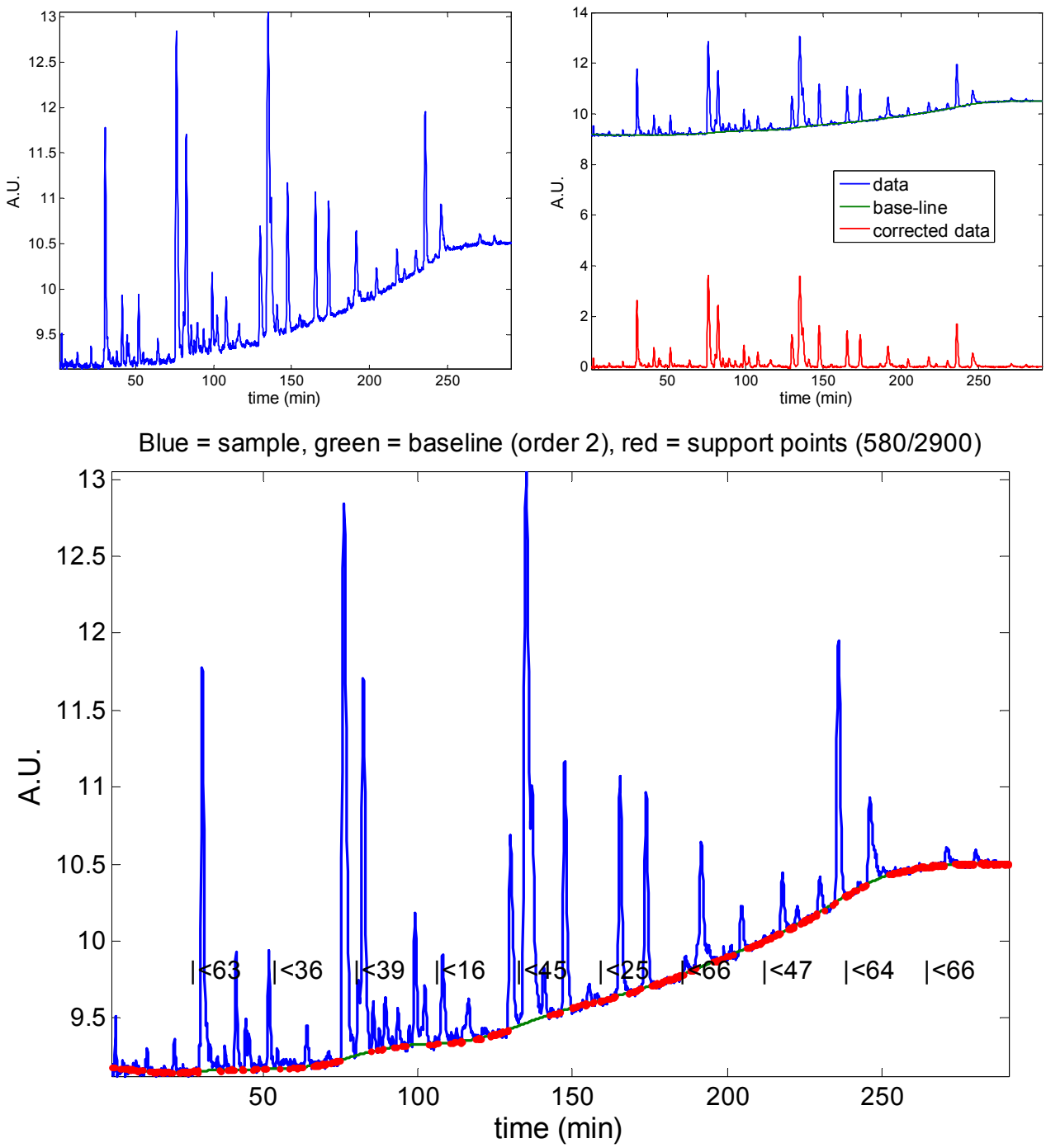


Figure 21. Illustration of baseline correction method by van den Berg (2008). TOP – left: Raw data with severe column bleed due to temperature raised during the chromatographic run. Right: baseline corrected signal and fitted baseline. BOTTOM: Detailed information about knot position, number of baseline points (i.e. support points) between knots and the amount of support points. The | indicates the knot position and the corresponding number the support points in the region to the left. The last region contains 113 support points.

This method operates by gradually eliminating the points in the signal furthest (distance) away from the fitted polynomial until the number of selected support points (baseline points) is reached. In the example, the algorithm eliminates points until 580 (20% of all time axis

points) are left and to these points (and the knots) the splines are fitted. To be successful, all points belonging to the analytical signal should be eliminated otherwise different parameters must be set. These parameters are the fraction of support points, the number of knots and the order of the splines fitted.

For the global method by Gan et al. (2006) the order of the polynomial is the only setting and as such does not require much user interaction. In the global method by Eilers (2004), asymmetric least squares fitting is used and this requires an input for the weights to put on points above or below the fitted polynomial and also a term determining the smoothness of the fitted polynomial. The algorithm is fast and uses only a few iterations to find a proper solution and as such proper input parameters are not critical. The method by van den Berg (2008) works in local regions. This requires 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 fit between the knots. For this method a lower order polynomial can often be used compared to the two other global methods. This means that fewer parameters must be calculated in each region, but the many local regions result in a method slightly slower than the global methods.

For simple chromatograms with perfectly resolved peaks and thus, many potential baseline points between peaks, all three methods will work well. For problematic baseline shapes the global solutions will have problems correcting local features, but in general temperature programs provides rather smooth baselines that can be fitted by relatively low order polynomials (in Figure 20 an order of seven was used). More complex chromatograms have less available baseline points and thus, removing the baseline becomes a more difficult task. But, these three methods offer a good potential even in the last mentioned case as was demonstrated by Gan et al. (2006).

4.2.2 Methods for multidimensional chromatographic data

Liang et al (1993) reported a method for detecting and correcting baseline offsets and drift in hyphenated chromatographic data. Their method works by constructing a principal component analysis model for the spectra from the baseline regions before and after the peak (zero-component regions), followed by a linear regression step to estimate a drifting or constant spectral background under the peak. If the spectral background showed slight variation during elution, an average spectral background was used. This method requires that zero-component regions can be found (or in other terms a chemical rank of zero – if one considers the background as a chemical signal then a rank of one is expected – see below, which can be assumed for regions with no eluting peaks and only noise). In complex chromatograms (several successive co-eluting peaks) these zero-component regions can be difficult to find. As will be shown in Chapter 5, zero-component regions (here assuming a

chemical rank of one) can be described and extracted by factor models due to their contribution to the overall measured analytical signal [PAPER IV].

4.3 Peak finding

Chromatographers go to great lengths to prepare, inject, and separate their samples, but they sometimes do not pay as much attention to the next step: peak detection and quantification. Identification and quantification of peaks can influence the quality of the results as much as or even more than sample preparation, separation, and detection.

It is often necessary to determine the position of peaks in the chromatograms along the elution time axis and the number of overlapping peaks in a given window. The simplest method would be to set an intensity threshold that the peak heights must exceed to be assigned as a valid real peak and not an uncertain peak highly influenced by noise. In this case the TIC chromatogram is an obvious candidate, but the success of this method depends on the success of the preceding alignment. With severe shifts, peaks can be assigned wrongly across samples. To prevent this, alignment should be done as explained earlier, but the use of the mass spectral information will also minimize the risk for false peak assignment. Having aligned peaks the peak detection method presented by Vivo-Truyols et al. (2005) or other derivative based approaches (Fellinger A., 1998b) have been shown to be efficient for well defined chromatographic peaks. Co-eluted peaks can also be handled by these methods and as such the only prerequisite is that data are properly aligned. Other approaches using the information from the mass spectral dimension have been proposed to either detect peaks or to match peaks from different chromatograms if no, improper or perfect alignment has been carried out [Dixon et al., 2006; Stein, 1999]. Dixon et al (2006) presented a method that in a semiautomatic way was capable of locating and matching peaks across several samples of severe complexity (i.e. hundreds or thousands of peaks).

For complex samples, several co-eluting peaks may occur in succession and an individual peak region/window might contain multiple peaks. To get an idea of the complexity needed for the subsequent curve resolution method, an initial rank determination in the found regions can be introduced. As will be mentioned in Chapter 5, Evolving Factor Analysis (EFA) is a useful tool in determining the local rank in a predefined elution time region. Determining the local rank over elution time regions before finding the location of peaks can also be used in situations where the peak region (start/stop) is difficult to set due to e.g. noise. Extending the window size to e.g. reach the baseline before and after the included peak(s) will result in a more complex peak system. However, this can be included in the subsequent curve resolution method provided that the complexity (i.e. chemical rank) of the system still can be considered as low-rank (further explained in the next chapter). As mentioned, some alignment techniques work by first finding specific landmarks and then aligning these landmarks across chromatograms. Landmarks are often well defined peaks

above a certain threshold, and as such a peak finding routine could already be included within the alignment procedure.

After proper peak alignment, baseline correction, peak location (window position and width) and local rank determination the data are ready for further inspection by multivariate curve resolution/deconvolution methods.

5 Modeling of chromatographic data

The next step in the data analysis of chromatographic data is at first to find accurate peak areas and/or informative fingerprints and secondly, to look for patterns that can describe the system. All this can be accomplished by chemometrics!

Multivariate data analysis has been applied for decades in order to extract information from multiple samples measured by several parameters/variables/attributes (i.e. wavelengths, elution times, chemical parameters, sensory attributes). For this to be possible the relevant variation must be readily available, meaning that instrumental artifacts must be removed. This is handled by alignment and baseline correction (and other techniques if needed – Chapter 2 and 3).

This chapter will include all aspects of analyzing chromatographic data using the multivariate modeling part in chemometrics (chemometrics can also be said to include other parts of mathematical methods for chemical data such as pre-processing). Here chemometrics will not only be two-way multivariate methods suitable for tables of peak areas or fingerprints but also multiway methods, where the advantage of including the additional spectral dimension (e.g. MS or UV) is significant. To be able to explain the multivariate principles (from now on this term includes both two-way and multiway methods unless otherwise stated in text) to a broad chromatographer/chemometrician audience, the principles will be explained using model terms and visualizations related to chromatographic data in general. The theory of the techniques will not be described comprehensively, as several text books are more suitable in this perspective; e.g. (Esbensen, 2002; Naes et al., 2002) for multivariate data analysis in general and more advanced methods in Smilde et al. (2004). Rather the usefulness for chromatographic data will be discussed and examples of bringing chromatography and advanced multivariate data analysis together will be given.

For chromatographic data, multivariate data analysis can be divided into two broad groups; 1) analysis of fingerprints or whole chromatographic profiles (continuous data) and 2) peak area estimation using factor models followed by multivariate analysis of peak areas (discrete data).

Notation

The notation and terminology to describe matrices (two-way arrays) and higher order arrays is adapted from Kiers (2000). Scalars are indicated with lower-case italics (e.g. x_{ijk}) and column vectors with bold lower-case characters (e.g. $\mathbf{y}$). T in superscript is the transpose operation (i.e. $\mathbf{x}^T$ is a row vector). Ordinary two-way arrays (matrices) are denoted $\mathbf{X}$ (bold-face) whereas higher order arrays are denoted $\underline{\mathbf{X}}$. The ijk^{th} element of a three-way array $\underline{\mathbf{X}}$ is

denoted x_{ijk} where the indices run as follows: $i = 1\ldots,I$; $j = 1,\ldots,J$; $k = 1,\ldots,K$. Three-way arrays will be denoted $\underline{\mathbf{X}}$ ($I \times J \times K$) where I is the number of samples, J the number of scans (elution time) and K the number of mass channels. Thus, the ijk^{th} element in $\underline{\mathbf{X}}$ corresponds to the mass detector response from sample i, measured at elution time j, at mass channel k. For multivariate models, bold capital letters are used to indicate scores, loading and other matrices despite the number of columns (i.e. number of components).

Terms for multivariate data analysis

This small section will describe some of the important terms when doing multivariate data analysis of chromatographic data. This will bring chromatographers and chemometricians on equal basis for the following sections.

(1) Mean centering and scaling
- In multivariate data analysis and especially when dealing with models that focuses on variability in data, it is standard to *mean center* the samples (subtracting the average chromatogram) to remove a common offset. This brings each variable to vary around zero.
- Discrete/non-continuous variables (e.g. peak heights and areas) will have different variance due to differences in concentration and response to concentration. This gives some variables relatively higher variance than others and as this may not reflect a higher importance, it is common to divide (*scale*) each variable by its standard deviation. In this way all variables get equal variance and enter the model on equal foot. Continuous variables (e.g. scans in the elution time dimension) are normally not scaled, as this will lower the importance of the peaks relative to the noisy baseline variables.

(2) Rank (mathematical and chemical)
- For chromatographic data the rank will always refer to the chemical rank and not the *mathematical rank*. The mathematical rank is the number of linearly independent rows or columns of the data matrix. For a real chromatographic data the mathematical rank would be equal to the minimum number of samples or variables (full rank) due to the instrumental noise in the data.
- One way to define the *chemical rank* of a data matrix is the number of linearly independent and chemically relevant factors. Thus, finding the chemical rank one has to separate chemical information from background and noise. In this thesis, the term rank will refer to the chemical rank of a data matrix or array and extracted features are said to be chemically meaningful if they follow the nature of chromatographic signals (one elution time and mass spectral profile per analyte).

(3) Peak systems and fingerprints (recaptured from Footnote 6 and 7).
- A *peak system* is here used as a description of a low-rank part of the data array; i.e. a part with contributions from only few chemical compounds. Each sample in a peak system is described by a two-dimensional matrix (e.g. GC-MS).
- This is in contradiction to the use of *fingerprints*, which refers to global parts of the data array, where one dimension has been collapsed (e.g. by summing over all variables in one dimension).

(4) Constraints
- With some multivariate models it is possible to constrain the solution to identify the model or to ensure that the model parameters make sense. E.g. *orthogonality* constraints are applied for identifying a PCA model, while *non-negativity* constraints are often applied in curve-resolution models because the underlying parameters are known not to be negative (e.g. for spectra or chromatograms). Also if a single peak system is modeled, unimodality can be imposed as it is expected that the profile can be described as a one peak/maximum (*unimodal*) curve. A constrained model will always fit data poorer but if the constrained model provides more interpretable results this can justify the decrease in fit.

(5) Component vs. factor and mode vs. dimension
- In multivariate methods these terms are often interchanged and used for the same. All the presented methods in the following are so-called factor models and when extracting information from a data matrix this is done using *factors*/*components*.
- The dimensionality of chromatographic data was discussed in Chapter 3. When dealing with multivariate models the use of both *dimension* and *mode* will be applied, e.g. the elution time dimension will also be referred to as the elution time mode.

5.1 Multivariate data analysis: Two-way

The workhorse and most basic multivariate analysis technique is Principal Component Analysis (PCA). PCA was originally developed by Pearson in 1901 [Pearson, 1901], though it is more often attributed to Hotellings work from 1933 (Hotelling, 1933) where he described and developed PCA to its present stage. Since then PCA has been used for several applications in different scientific disciplines. Besides this, PCA is implemented in all multivariate data analysis software packages and also in some instrument software. In the chromatographic discipline PCA has also found its use and this has been formulated in several reviews [Brakstad, 1995; Brereton, 1995; Christensen and Tomasi, 2007; de Juan and Tauler, 2007; Duarte and Capelo, 2006; Liang et al., 2006; Peris, 1996].

PCA is a bilinear model that searches for common patterns in a data table (two-way array) by establishing new directions in the original data cloud; so-called latent variables or

loadings, which are constructed as linear combinations of the original variables [Wold et al., 1987]. The first new direction is found so that the maximum variance in the original data cloud is explained. For the first direction, each sample (from its original position) can be projected onto this, providing a *score* value. These score values then describe the amount of the latent variable/loading found in each sample. A set of a score and loading vector constitutes what is denoted *a principal component*. 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. The variance left in the data (unexplained variance) is usually related to unsystematic variation or noise and is termed the residuals. To illustrate this from chromatography a simple one-peak system is shown in Figure 22.

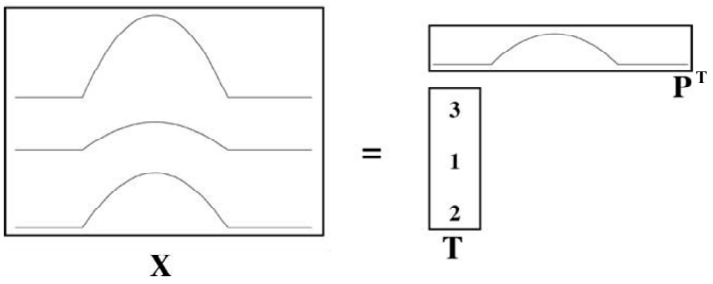


Figure 22. Illustration of Principal Component Analysis (PCA) using a simple chromatographic one-peak system for three samples. **X** is the original data, **P** the loading vector (latent variable/common profile) and **T** the score vector holding the amount of the common profile. Modified from Skov and Bro (2005). No noise is present in the data and thus one principal component ($R = 1$) will explain all variance (rank one system); i.e. no residuals.

As seen, this is a simple rank one system, as the only difference between the peaks in the three samples is the peak height (same chromatographic profile). The PCA model captures the maximum variation, which follows the peak profile and thus, the first loading resembles the original data, as expected for this simple example. The score value is then simply a measure of the magnitude of the peak profile and can be used as a direct measure of relative peak area or concentration. The latter is a consequence of the model being bilinear meaning that twice as high concentration (peak area) gives twice as high a score value (assuming no noise and baseline present and similar peak shape regardless of the concentration).

For real chromatographic data, noise and interfering compounds are always present. In this case a more comprehensive PCA model is needed including a noise term as well. This more detailed version of PCA has been indicated in Equation 7. The raw data (always mean centered) can be expressed as a sum of the modeled part; the outer product of the **T** (scores) and **P** (loadings) matrices plus the not modeled part, the residual matrix, **E**.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E} \quad (7)$$

The number of principal components extracted determines the number of columns in both **T** and **P** and as such **X** could also be described as the sum of vector products ($\mathbf{t}_r \times \mathbf{p}_r^T$), where index $r = 1,\ldots,R$ with R being the number of components in the model, plus the residual matrix, **E**. One important aspect of the factor models shown here is that the loadings are normalized to a length of one by common convention in chemometrics. In this way the 'amount' of this new/latent direction (the position projected onto this vector of length one) will be directly translated and held in the score values. In PCA, the extraction of principal components is done using the orthogonality constraint. This means that different variability is explained in the successive components. This also means that an R+1 component model contains the solution to an R-component model plus one additional component. E.g. the two first components of a three-component model are identical to the components of a two-component model. Thus, PCA is said to be a *nested* model.

Basically, PCA can be used in several ways for chromatographic data; 1) Peak systems for individual samples to find the complexity (number of analytes) of the system (data: elution times × mass channels), 2) Local fingerprints (only one peak system is investigated) to classify samples based on only the content of only few analytes (data: samples × elution times or mass channels) and 3) Global fingerprints to classify or describe the whole set of samples based on the overall profile (data: samples × elution times or mass channels). For the latter two, one dimension must be collapsed and as explained, for GC-MS data this is done by summing data over variables in the 'unimportant' dimension (see Chapter 3 for more on data structure and dimensionality).

To illustrate the multivariate techniques for peak deconvolution or local fingerprints, a simple mixture system of two aroma compounds found in cheese (2-methyl butanal and 3-methyl butanal) will be used throughout this chapter. In Figure 23 this simple system is shown for five mixtures in triplicates (more information about the mixture system can be found in the figure legend and PAPER II).

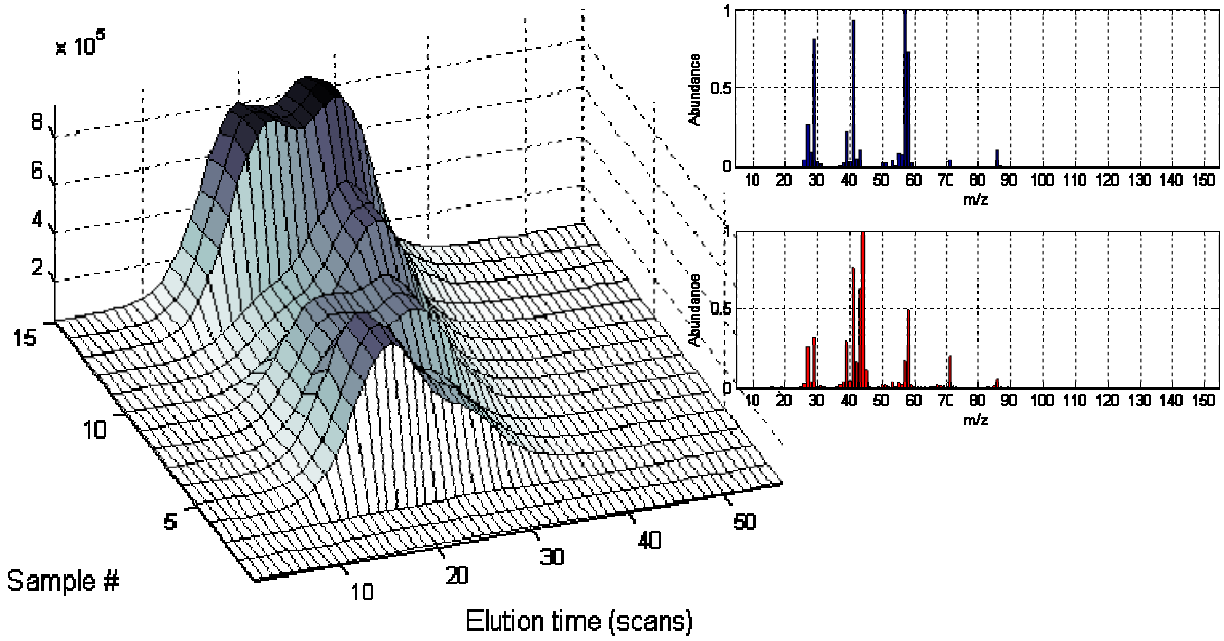


Figure 23. Left: Example of TIC chromatograms for a mixture of two important aroma compounds found in cheese (A: 2-methyl butanal and B: 3-methyl butanal). Here analyzed as a mixture in heptane on a polar 30 m long DB-Wax capillary column with an internal diameter of 0.25 mm, and with a 0.25 µm film (thickness). Sample 1-3: 0 ppm of both, sample 4-6: 0.75 ppm of A and 0.25 ppm of B, sample 7-9: 0.50 ppm of both, sample 10-12: 0.25 ppm of A and 0.75 ppm of B and 13-15: 1.00 ppm of both. Right: EI Mass spectra of the two aroma compounds – top: 2-methyl butanal and bottom: 3-methyl butanal. Figure from PAPER II.

For the global fingerprints a larger data set of red wine samples from two different geographical regions will be used and explained further in section 5.1.3.

5.1.1 Single peak system – one sample (landscape)

In Figure 24 a one-sample system analyzed with PCA, but this time including the mass spectral information is shown. This enables the model to look for differences over the elution time regions found in different mass channels. It is expected that some mass channels will contain the first peak profile whereas others contain the second peak profile. For PCA this additional information is providing a better understanding of the system, but the extracted features are still not chemically meaningful as the extracted elution time and mass spectra are not related to single-peak features. One can say that PCA extracts components in a *mathematically* meaningful way (describes directions of maximum variation that is not related between extracted components – orthogonality constraint). This is in contrast to other more dedicated factor models that are capable of extracting *chemically* meaningful information. This will be treated in details in section 5.2.

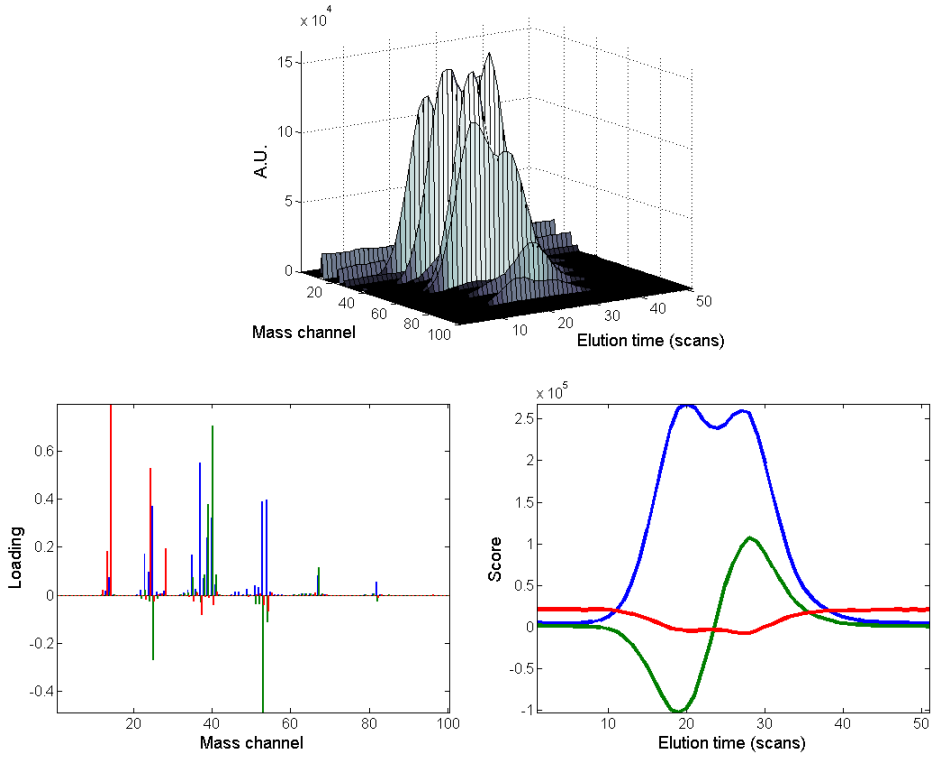


Figure 24. Two-dimensional signal for one sample (e.g. GC-MS or HPLC-DAD) – on of the green samples in top left plot of Figure 25. TOP: Raw data. BOTTOM – left: Loadings and right: Scores. Order of principal components (PC) is blue (PC1), green (PC2) and red (PC3).

The first approach of PCA to chromatographic data actually dates back to 1972, where Macnaughtan and coworkers developed an algorithm that under certain restrictions resulted in a proper description of a chromatographic peak system [Macnaughtan et al., 1972]. Even though PCA-like models can be applied for peak systems, the major use of PCA has been to get an idea of the complexity of a co-eluting system – i.e. number of overlapping peaks in the elution window of interest. Stated differently, the target of PCA is to guess (i.e. find) the number of variability sources up to the noise in the data. For the latter issue, Evolving Factor Analysis, EFA, can also be applied. EFA offers similar information as PCA and, furthermore, EFA provides an intuitive idea of the elution profile for overlapped peaks. EFA does rank determination over a predefined elution time region for estimation of significant contribution to the local rank of the system [Maeder, 1987; Maeder and Zilian, 1988; Roach and Guilhaus, 1992]. Both PCA and EFA are often used prior to more dedicated curve resolution/deconvolution methods.

5.1.2 Local fingerprints – more samples

For local fingerprints we seek a pattern that can describe the peak or peaks in a small region. With just one peak, a similar decomposition of the data as the one shown in Figure 22 could

be achieved including the error term holding the instrumental noise. However, with two or more peaks it is not straightforward to extract the individual peak profiles. In Figure 25 one example of this has been shown, where a three-component PCA model has been calculated for the mixture system. In the example, the loading plot shows that no chemically meaningful elution time profile for any of the two aroma compounds is obtained. Despite this, the power of PCA can be observed in the score plot. PCA is able to cluster the five mixtures as they have a different pattern captured in the first two principal components. The first component describes the overall pattern (a kind of weighted average of the peak profiles) and the second component then corrects for the observed difference (in raw data) from this average.

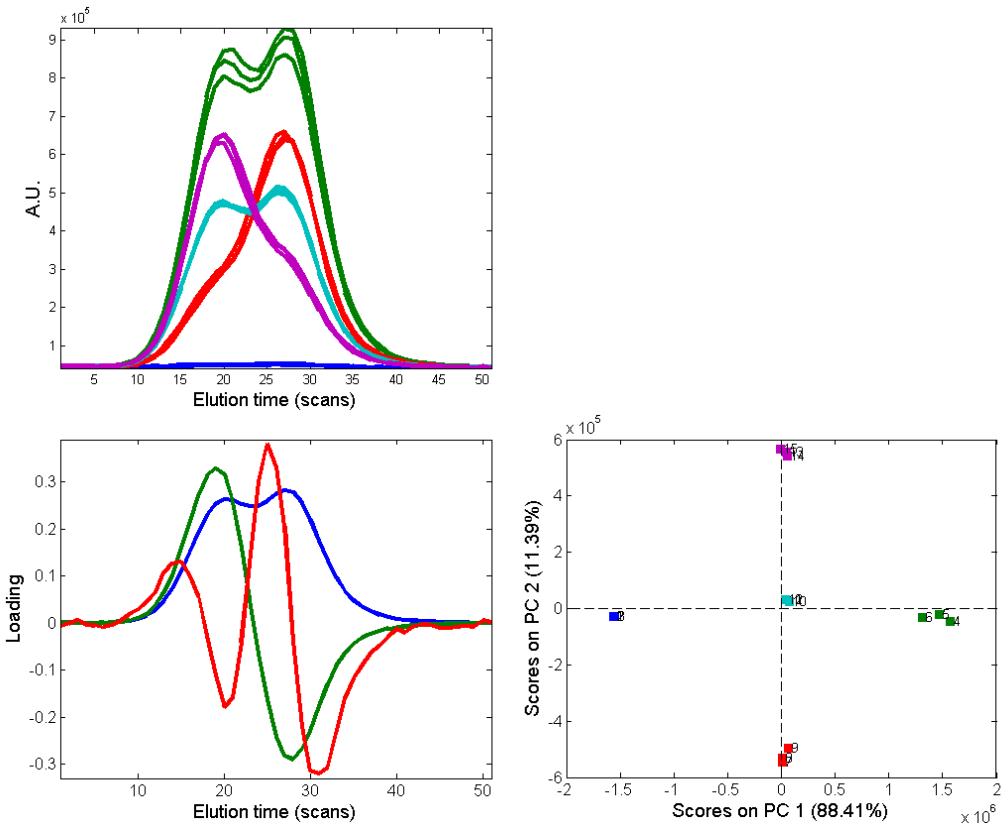


Figure 25. Illustration of the inadequacy of PCA to deconvolute overlapping peaks into real chemical meaningful model components. TOP: One-dimensional signals (e.g. summed MS – TIC chromatogram, GC-FID or single-channel MS). BOTTOM – left: Loadings and right: Scores. Order of principal components (PC) is blue (PC1), green (PC2) and red (PC3).

The PCA model reveals that three mixtures have the same profile as the one described by the first loading (green, cyan and blue curves – these samples have the same content of the two aroma compounds). The direction spanned by these three mixtures describes the main variation found in the data (the first loading) and the amount of this latent variable is given by their score values. The second loading then describes the mixtures with a different profile by describing their deviation from the first overall profile. As this loading crosses zero in the

middle of the two peak maxima, this compensates for the uneven content of the two aroma compounds in these mixtures (purple and red curves). Even though only two components were expected to vary in the mixture system extracting the third principal component does not change the interpretation of the PCA model (as the model is nested).

5.1.3 Global fingerprints – more samples

As observed from Figure 25 PCA was capable of describing the samples efficiently (clustering of known mixtures in the score plot) without focusing on providing interpretable loadings. This brings us to the next level in the use if PCA; the modeling of large data tables to find patterns obtained from chromatographic one-dimensional profiles/fingerprints. A simple example can be given for two classes of samples such as wines from different regions. Provided that proper sampling has been done, these two classes should have a distinct fingerprint that characterizes the geographical region (assuming grape related, environmental and growth factors to be the same). This is visualized in Figure 26.

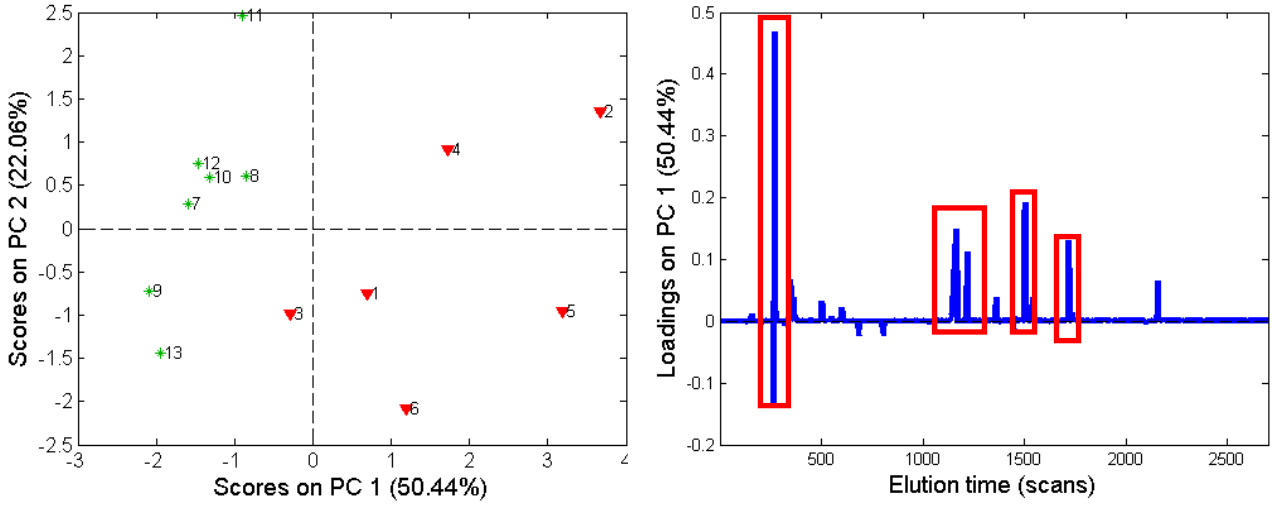


Figure 26. Example of a PCA model on thirteen wine samples from two different geographic regions. LEFT: Score plot of wine samples colored after class relationship and RIGHT: Loading plot for the first principal component that separates the two classes in the score plot. NOTE: Samples have been centered and variables have not been scaled as is the general approach for continuous/spectral data. For more on centering and scaling of both two-way and multi-way data see Bro and Smilde (2003).

In Figure 26 the two geographical regions are separated along the direction of the first loading vector, meaning that the pattern described here is found to be most significant in wines to the right in the scores plot. Peaks with high loadings in the first principal component are important and these can be further studied by local peak deconvolution or integration as described in the previous section and more detailed in the next section. The second main pattern (loadings in the second principal component) is more difficult to interpret without knowing more about the data, but aspects such as the ones assumed to be

constant or other factors could be sources to this variability. For data with well-known classes, more dedicated supervised methods can be used. Local regions in chromatograms that are responsible for the class separation can be found using Interval PLS (iPLS) (Norgaard et al., 2000) and for testing the class relationship, discriminant PLS or Soft Independent Modeling of Class Analogy (SIMCA) (Wold and Sjostrom, 1977) have proven to be advantageous.

5.2 Multivariate data analysis: Multi-way

For hyphenated methods such as combining a chromatographic separation system with a spectroscopic detector, a data table/landscape for a single sample is obtained. In the previous section this was handled by PCA by either summing the spectral dimension or by analyzing one sample at a time. Both PCA approaches are suitable for an exploratory analysis of patterns in the data or for evaluating an overlapping peak system. However, PCA was inadequate for finding direct chemically meaningful information, but this can be dealt with using more advanced curve-resolution methods or factor models such as MCR-ALS (Tauler, 1995), PARAFAC (Bro, 1997; Carroll and Chang, 1970; Harshman, 1970) and PARAFAC2 (Bro et al., 1999; Harshman, 1972; Kiers et al., 1999; Wise et al., 2001). Recent reviews recapture the use of factor models in general for chromatographic data (de Juan and Tauler, 2007) and calibration models based on chromatographic data [Escandar et al., 2007; Ortiz and Sarabia, 2007].

Several of these methods have been applied to GC-MS data (see references mentioned in the following), but in this thesis the focus will be on the PARAFAC and PARAFAC2 models. A more detailed description will be provided later, but here some initial statements and motivations regarding the use of PARAFAC and PARAFAC2 will be given for GC-MS data.

1) Multiway data should be modeled by multiway models!
2) The internal data structure is kept intact, as there is no need to sum/collapse certain dimensions (e.g. summing the mass dimension to obtain TIC chromatograms), which can result in pre-modeling/initial loss of information.
3) In theory GC-MS data is trilinear as each analyte should have its own mass spectral and elution time profile and being able to find these, means that the amount of analyte (relative concentration) is linear with respect to these profiles. However, shift in elution time across samples means that data will often not be trilinear. Shown earlier, shifted profiles can be aligned and trilinear models (e.g. PARAFAC) can be applied. Without alignment, models that handle these shifted profiles can also be applied (e.g. PARAFAC2).
4) Both PARAFAC and PARAFAC2 provide a unique solution to a given low-rank problem, which for chromatographic data is normally a system of few

overlapping/co-eluting peaks. The uniqueness is due to models being able to find unique mass spectral and elution time profiles for the individual analytes in the system investigated that will be chemically meaningful if the proper number of factors is included in the model. Unique profiles do not indicate that selective[8] variables, e.g. selective mass channels, are needed (Sinha et al., 2004a; Sinha et al., 2004b). PARAFAC and PARAFAC2 provide chemically meaningful solutions with no selective mass channels provided that the ratios of the intensities of the individual mass channels are different (i.e. different patterns). Selectivity is often required using commercial chromatographic software and these programs have difficulties when no selective ions are present (e.g. for isomers).

5) Scores and loadings are obtained as for PCA that can be used for visualization. Here, however, these parameters have a chemical interpretation.
6) Additional loading plot(s) for the extra dimension(s), which can be used to identify the analyte of interest (e.g. compared to modeling the TIC chromatogram with PCA).

PARAFAC can be regarded as a natural but constrained extension[9] of PCA in N dimensions (although an extension, there are differences and these will be discussed later). To illustrate how this model handles data compared to PCA, the modeling of data visualized in Figure 22 and Figure 25 using PCA are shown in Figure 27 using PARAFAC. Here no modes must be collapsed and sample, elution time and mass spectral dimensions are kept intact.

[8]Selective variables are here defined as variables which can be used to select between the compounds investigated. E.g. ions/fragments present in the mass spectrum of *only one* of the overlapping compounds.
[9] Sometimes the natural extension of PCA to multidimensional data with respect to many of the properties of PCA is said to be the Tucker3 model (see Section 5.3).

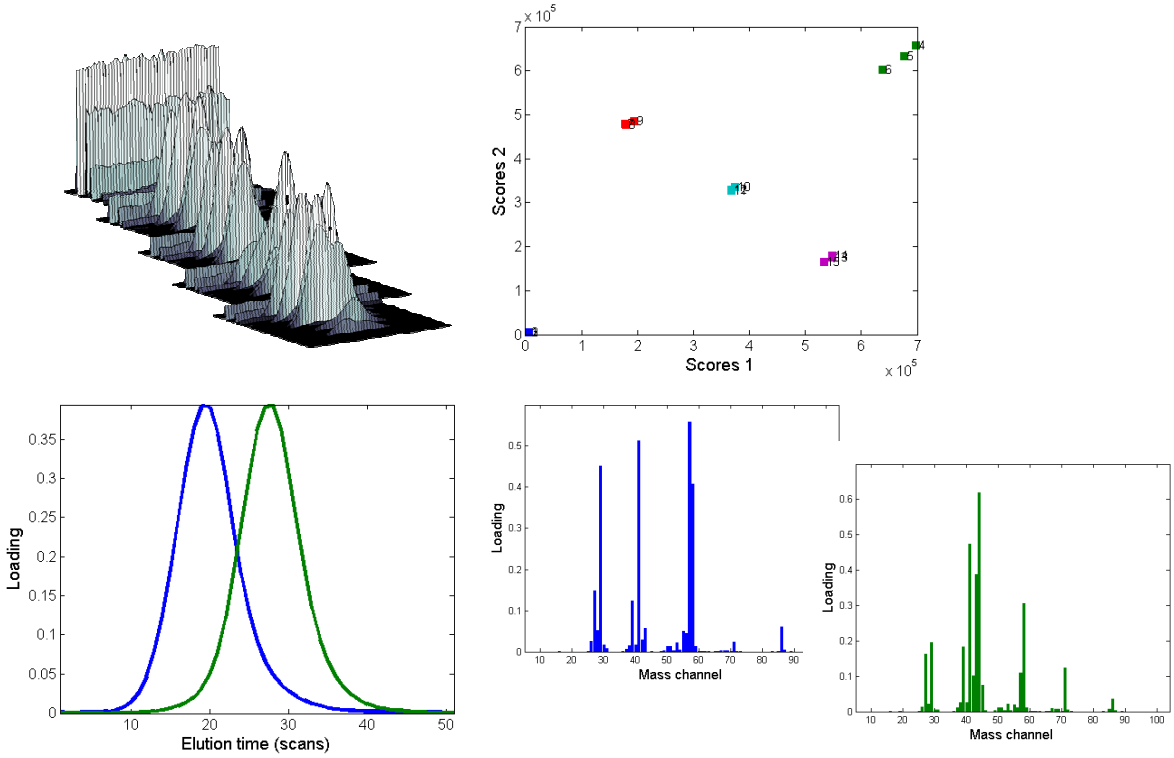


Figure 27. PARAFAC model of data shown in Figure 25 without summing any dimensions and analyzing all fifteen samples in one model. TOP LEFT: Raw data showing only one replicate of the five mixtures, RIGHT: Scores plot (score 1 vs. score 2). BOTTOM LEFT: Loadings for elution time mode and RIGHT: Loadings for mass spectral mode (notice the similarity to the pure EI pure mass spectra of the two aroma compounds visualized in Figure 23).

Comparing the PARAFAC (Figure 27) and PCA (Figure 25) models, the additional mode in PARAFAC provides an overall much better description of the data. As mentioned, PARAFAC can model a low-rank system and provide unique loadings for the chromatographic and mass spectral modes (notice the great similarity of the extracted mass spectral loadings and the pure mass spectra of the two aroma compounds from Figure 23). The two analytes are now being described by one factor each (blue and green curve) and not by more factors, as was the case for PCA. A similar score plot was obtained by PCA, but this just indicates that PCA can indeed model the system, but instead of focusing on chemistry in PCA the deconvolution is mathematically orientated. For PARAFAC, the splitting of data into factors holding single analyte contributions means that the scores and loadings are useable for e.g. calibration models, identification purposes and more. E.g. knowing the concentration of just one of the fifteen samples, the remaining concentrations for both analytes in all mixtures can be estimated based on the extracted score values [Bro, 2003]. Comparing PCA and PARAFAC might seem unfair due to the different information in the data described. However, this just illustrates the need for keeping the data intact and use multiway models for multiway data whenever possible!

The decomposition of data by PARAFAC into real analyte contributions resembles an enhanced chromatographic separation obtained by changing the column or instrument settings to resolve overlapping peaks on the elution time axis. This is the reason why PARAFAC (and PARAFAC2) is often described as performing *mathematical chromatography*.

5.2.1 PARAFAC

Multidimensional chromatographic measurements provides data of more than the traditional two dimensions (samples × variables) and this gives new possibilities with regard to the information that can be extracted. There are methods that make specific use of the so-called second order or multi-way structure of such data; a feature that can be used to quantify analytes in the presence of unknown interfering chemical compounds (Booksh and Kowalski, 1994; Boque and Ferre, 2004; Bro, 2003; Comas et al., 2004; de Juan and Tauler, 2007; Escandar et al., 2007; Ortiz and Sarabia, 2007; Rinnan et al., 2007), which would otherwise require a more comprehensive calibration.

PARAFAC was shortly introduced in the previous section where its performance on GC-MS data was visualized. The principles of PARAllel FACtor analysis (PARAFAC) was originally proposed in 1970 independently by Harshman (1970) and Carroll and Chang (1970) (the latter authors named the technique CANDECOMP) and has been used to model multiway chromatographic data to get both qualitative and quantitative information (Bro, 1997; Bylund et al., 2002; Hoggard and Synovec, 2007; Johnson et al., 2004) even when several peaks are overlapped. In PCA underlying features were extracted successively using orthogonality constraints (model is nested), but this strict mathematical constraint is not used in PARAFAC. Here the extracted components are allowed to be related to some extent as long as they differ enough in the elution time and mass spectral profiles to be identified as individual contributions to the overall signal. PARAFAC is also not nested (the first component of a two factor model does not hold the same information as a one factor model) and only when the proper number of factors are extracted/determined, the model will provide unique elution time and mass spectral loadings/profiles, which are at the same time chemically meaningful (the PARAFAC model will always provide a unique solution, but this solution is only chemically meaningful using the proper number of factors).

In theory, PARAFAC requires *low-rank* trilinear data (Bro et al., 2001 describes this issue for PARAFAC in comparison to more exploratory factor models such as PCA and Tucker3). Trilinearity can be viewed as an extension of Lambert Beer's law to second-order data. This amounts to assuming that the measured peak is the sum of the individual peaks of each analyte and that the elution profile and the mass spectrum of one analyte are proportional in all the samples [Comas et al., 2004].

To explain PARAFAC in terms of data arrays the model has been depicted in Figure 28 and Figure 29 and Equation 8 and 9.

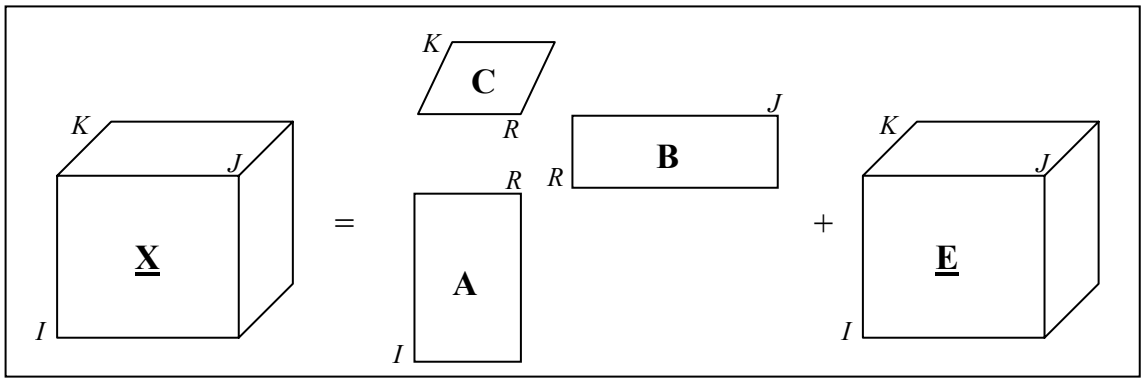


Figure 28. Visualization of PARAFAC model for GC-MS data of: SAMPLE × ELUTION TIME × MASS CHANNEL.

$$\mathbf{X}_i = \mathbf{B}\mathbf{D}_i\mathbf{C}^{\mathrm{T}} + \mathbf{E}_i \qquad i = 1,\ldots,I \tag{8}$$

where $\mathbf{X}_i$ is the ith frontal slab (sample mode) of the three-way array, $\mathbf{D}_i$ is a diagonal matrix holding the ith row of **A** in its diagonal and $\mathbf{E}_i$ residuals. **C** and **B** are the loading (column) matrices for the elution time and mass spectral mode, respectively. As for PCA, the loading vectors in **B** and **C** are normalized to a length of one for each component extracted and thus, the amount in each sample of the common elution time and mass spectral profiles can be found in the score value.

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{9}$$

where R is the number of factors included in the model and the indices i, j and k as explained earlier.

Following the PARAFAC model description of Figure 28 and Equation 9 a one sample mixture of the two aroma compounds can be described according to Lambert Beer's law terms as individual independent contributions to the overall mixed signal (Figure 29). Here three components have been extracted each described by a concentration term (a_{11},a_{12} and a_{13}) multiplied by the unique elution time and mass spectral profile. The sum of these individual contributions (plus the residuals) defines the overall mixed signal. The three components extracted are two analytes (aroma compounds) plus the background signal. Notice the resemblance to Equation 6. Here the analyte signal is simply split in to two terms.

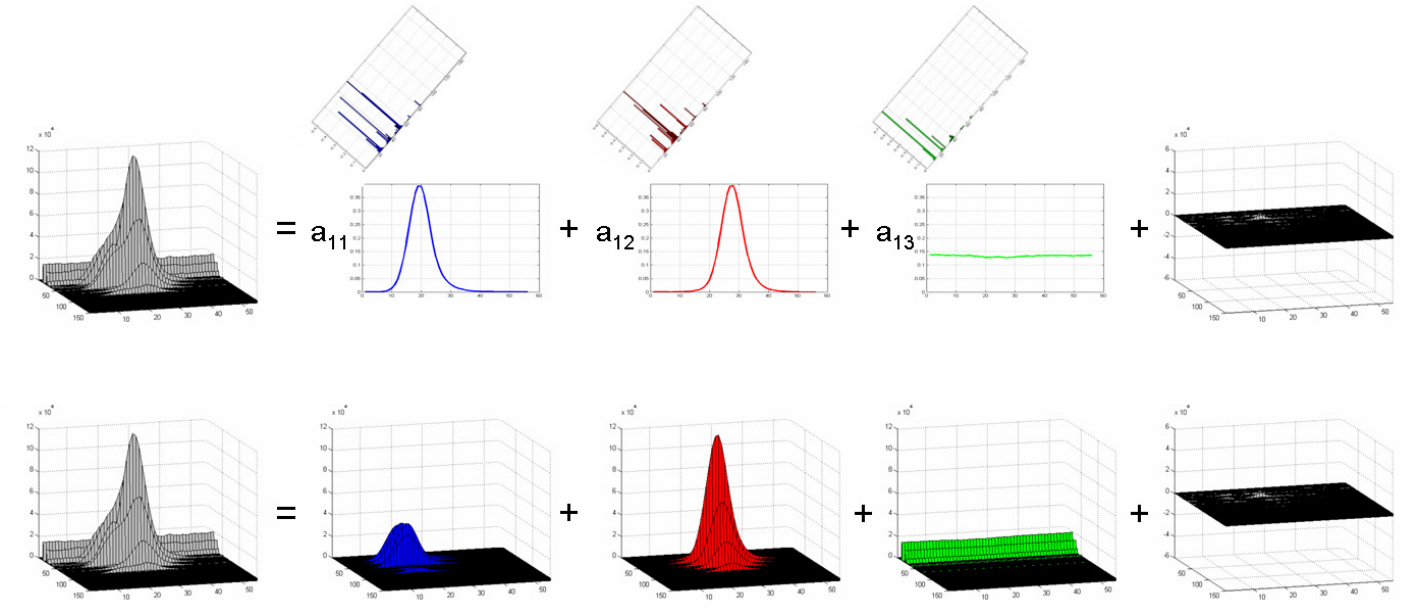


Figure 29. PARAFAC decomposition (using 3 factors) of the mixture of the two aroma compounds shown in Figure 23. For visualization only sample eleven (A: 0.25 ppm and B: 0.75 ppm) is shown here. Top: PARAFAC model shown with loading matrices for the pure chromatographic and mass spectral profiles for the two analytes and baseline found in sample eleven. Bottom: PARAFAC model shown with multiplied loadings. The unexplained part, the residuals, is shown to the outer right. Notice that the magnitude of the ordinate axes for the three factors (multiplies loadings) and the residuals are the same to indicate the model performance. Figure from PAPER II.

Having the deconvolution of the mixture experiment fresh in memory, Figure 28 shows that a three-way data array can be decomposed into one score and two loading vectors ($R = 1$) or matrices ($R > 1$) holding structured information and a residual array holding noise or unstructured information. One can see that collapsing the third mode (here mode **C**) by e.g. summing the mass spectral information the PARAFAC model is closely related to the PCA model discussed earlier. PARAFAC assumes invariant elution time and mass spectral profiles across the samples and as such, one prerequisite is that data are aligned prior to modeling. Another important thing besides synchronized profiles is that peak shapes must be similar for individual peaks across samples. Otherwise the obtained loading for elution time mode for an individual peak system would be a kind of average of the peaks shapes across samples. When reconstructing the peaks or using the score values for subsequent quantifications studies this would introduce errors. Peak shape changes can occur when one of the elution mechanisms in the column are overloaded – e.g. peak tailing due to too high concentration of a specific analyte [Dolan, 2002; Dolan, 2003].

To summarize, two things affect the low-rank trilinearity assumption intrinsic for PARAFAC; 1) the *degree of shifts* along the elution time axis across samples and 2) the *change in peak shape* for individual peaks across samples. The first aspect can be handled by

the alignment methods presented in Chapter 4, but the solution how to deal with changes in peak shape is less obvious. These two things can be handled efficiently by a more relaxed (less constrained) version of PARAFAC; PARAFAC2 described in the next section.

5.2.2 PARAFAC2

To handle both unsystematic shifts along the elution time axis and changes in peak shapes, a less (and different) constrained version of PARAFAC can be applied to low-rank peak systems. This model, called PARAFAC2, allows the natural structure of chromatographic data (e.g. shift peaks) to be present in the data while still providing a meaningfully unique solution. For PARAFAC2 each elution time profile (where the shift is located) is allowed to change both in pattern/structure and also in length provided that the cross product of the elution time profile is kept constant for all samples (explained below and in Figure 31 and Table 4). PARAFAC2 has been applied for several applications in diverse scientific fields from sensor based data (Skov and Bro, 2005), over flavor profiling (Ovejero-Lopez et al., 2005) and time-intensity studies (Reinbach et al., 2007) to kinetic data (Cueva et al., 2001). The reader is referred to the paper by Amigo et al. (PAPER IV) for a more complete list of PARAFAC2 applications.

The PARAFA2 model has been visualized in Figure 30 and Equation 10.

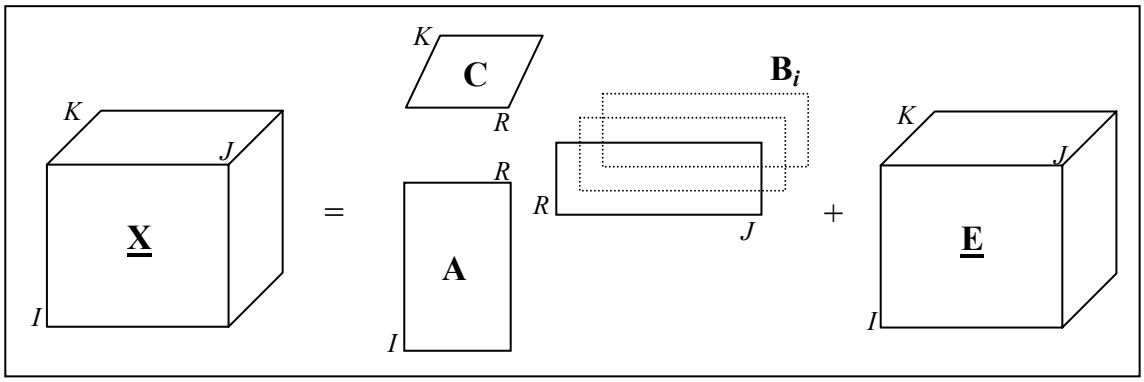


Figure 30. Visualization of PARAFAC2 model for GC-MS data of: SAMPLE × ELUTION TIME × MASS CHANNEL. Technical note: The presented order of modes follows the way GC-MS data are obtained and not how these data should be analyzed by e.g. in the PLS Toolbox by [Wise et al., 2006]. In contrast to PARAFAC, the order must be permuted to have samples in the last mode and the shifted profiles in the first mode.

$$\mathbf{X}_i = \mathbf{B}_i\mathbf{D}_i\mathbf{C}^{\mathrm{T}} + \mathbf{E}_i = (\mathbf{P}_i\mathbf{H})\mathbf{D}_i\mathbf{C}^{\mathrm{T}} + \mathbf{E}_i \qquad i = 1,\ldots,I \tag{10}$$

where $\mathbf{X}_i$ is the ith frontal slab of the three-way array, $\mathbf{D}_i$ is a diagonal matrix holding the ith row of $\mathbf{A}$ in its diagonal and $\mathbf{E}_i$ residuals. $\mathbf{C}$ is the loading matrix for the mass spectral mode and $\mathbf{B}_i$ the elution time loading matrix for the ith slab of $\underline{\mathbf{X}}$ modeled as $\mathbf{P}_i\mathbf{H}$. $\mathbf{P}_i$ is of size

($I \times R$) and **H** of size ($R \times R$). $\mathbf{P}_i$ and **H** have no direct chemical or physical interpretation, but as shown their product will be an estimate of the elution time profiles, $\mathbf{B}_i$.

As mentioned in the legend of Figure 30 the algorithms available for calculating PARAFAC2 models require that the dimensions are rearranged (permuted) to have samples as last mode and shifted profiles in the first mode [Wise et al., 2006]. In contrast to PCA and PARAFAC, the vectors in the sample mode are normalized to a length of one and the concentration information is then kept in the individual elution time loadings ($\mathbf{B}_i$). To include the concentration information in the elution time loadings, the individual $\mathbf{B}_i$ loadings could be multiplied with its corresponding scores $\mathbf{D}_i$. This is needed to make the comparison with the real elution time profiles straightforward (see Figure 36 for an example of these two situations). If peak areas in multiple peak systems should be compared, PCA and PARAFAC allow a direct comparison of score values from different models (as loadings are normalized in these models). But, to be able to do the same with PARAFAC2, the individual elution time loadings $\mathbf{B}_i$ must be rescaled to a length of one and at the same time rescaling the scores (sample loadings) so that the reconstructed peak is intact.

5.2.3 Uniqueness – *in short*

The PARAFAC solution cannot be rotated without a loss of fit and hence only one best-fit solution (unique) is possible for a model using R factors. For the PARAFAC2 model to be valid and to retain uniqueness, all the cross-products matrices of the $\mathbf{B}_i$ matrices are constrained to be constant for all i. This can be formulated as [Kiers et al., 1999]:

$$\mathbf{B}_i^{\mathrm{T}}\mathbf{B}_i = \ldots = \mathbf{B}_I^{\mathrm{T}}\mathbf{B}_I \qquad i = 1,\ldots,I \tag{11}$$

This means that for every sample i, a set of profiles $\mathbf{B}_i$ (e.g. elution time profile Bro et al., 1999) is estimated under the constraint that the cross-products of the loading matrices are identical. More on uniqueness for PARAFAC and PARAFAC2 can be found in the monograph by Bro (1998) and in a recent textbook by Smilde et al. (2004).

In Table 4 and Figure 31, three simulated GC-MS peaks are given (**X**), (**Y**) and (**Z**). They are the same peak shifted by one and two positions along the elution time axis, respectively. The inner-products ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ and $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$) yield identical values.

Table 4. Simulated GC-MS data for three situations (elution time × mass channel) shifted along the elution time axis. The inner product is indicated below the three matrices (modified from van Mispelaar et al., 2003). Notice how the peak maximum (bold letters) shifts along the elution time axis.

Situation X					Situation Y					Situation Z				
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	3	6	3	0	0	0	2	0	0	0	0	0	0	0
0	5	8	5	0	0	3	6	3	0	0	0	2	0	0
0	7	**10**	7	0	0	5	8	5	0	0	3	6	3	0
0	5	7	5	0	0	7	**10**	7	0	0	5	8	5	0
0	0	4	0	0	0	5	7	5	0	0	7	**10**	7	0
0	0	0	0	0	0	0	4	0	0	0	5	7	5	0
0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
Inner product														
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	108	163	108	0	0	108	163	108	0	0	108	163	108	0
0	163	269	163	0	0	163	269	163	0	0	163	269	163	0
0	108	163	108	0	0	108	163	108	0	0	108	163	108	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

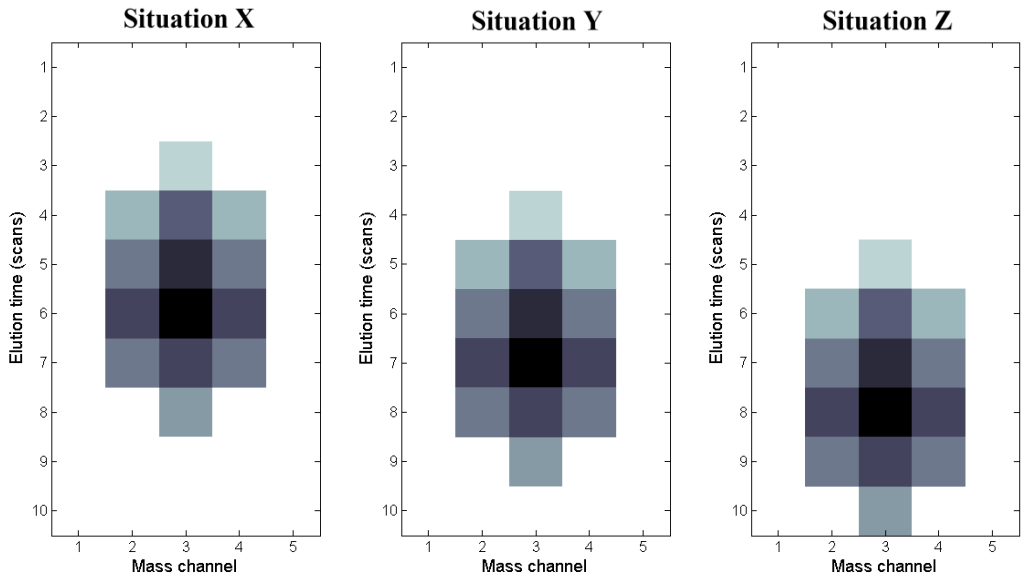


Figure 31. Visual depiction of the three shifted but otherwise identical GC-MS peaks (landscapes) indicated by letters (**X**, **Y** and **Z**) in Table 4. Normally more mass channels are included, but this situation could be a case of only looking at the five most intense mass channels for the given peak.

It is seen that when the peaks are shifted (similar can be shown for shape changes across samples) the same inner product matrix is achieved. This principle is similar for the estimated elution profiles $\mathbf{B}_i$; $\mathbf{B}_i^{\mathrm{T}}\mathbf{B}_i$ must be the same across samples. In a one factor

PARAFAC2 model (rank one), the size of $\mathbf{B}_i$ becomes $J\times1$ and the cross product 1×1. If more factors are extracted the size of the cross product becomes $R\times R$ and as seen the length of $\mathbf{B}_i$ is allowed to change and still provide similar cross products. In the example above the first row of situation **X** could be removed (changing the length of the elution time profile but preserving the peak profile) and still provide the same inner product.

For better understanding of PARAFAC and PARAFAC2 similarities and dissimilarities a small example of a two-analyte system is shown in Figure 32 with shifts introduced in the elution time dimension.

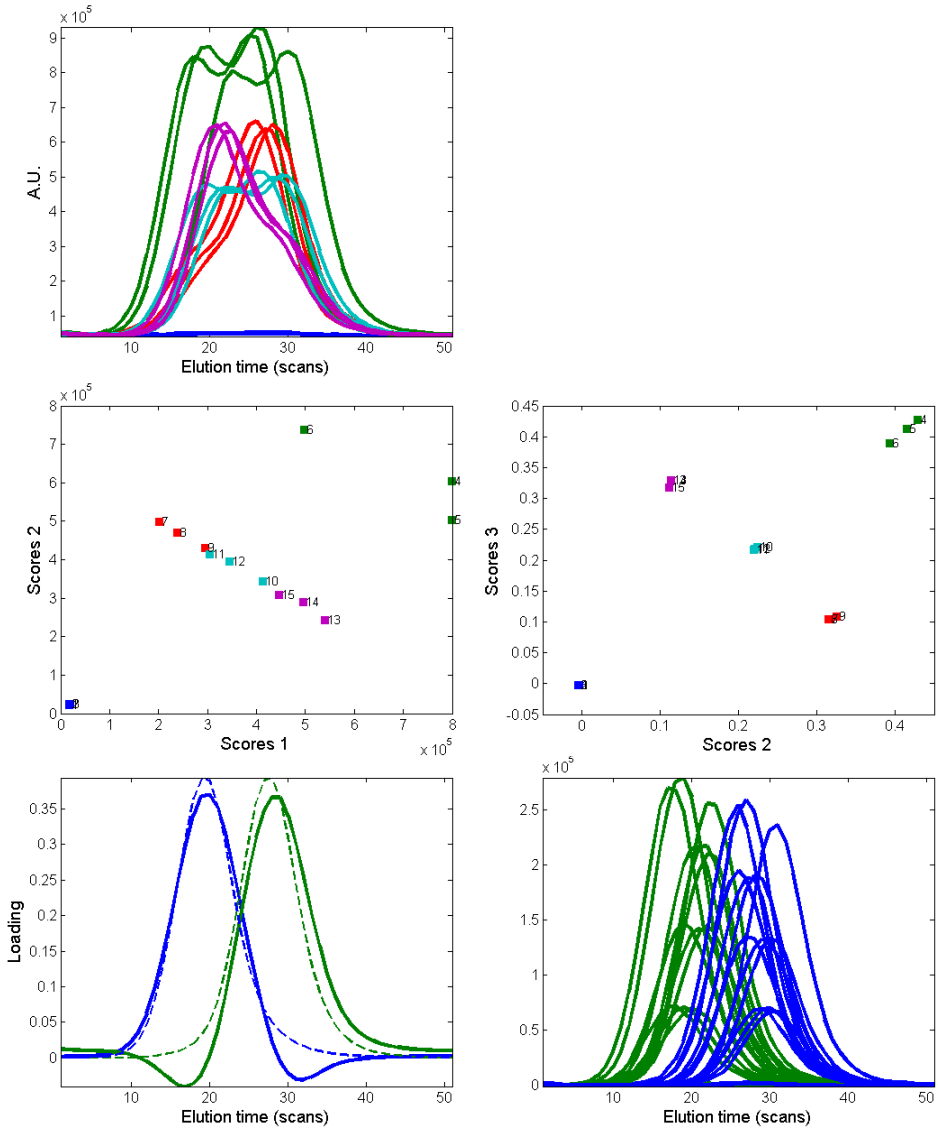


Figure 32. Illustration of PARAFAC and PARAFAC2 for shifted data. TOP LEFT: Raw data as shown in Figure 25, but now randomly shifted an integer between zero and four in both directions. Data are shown as TIC chromatograms to enhance the depiction of the shifts. LEFT: PARAFAC scores and elution time loadings. Note that the loadings for the not shifted PARAFAC model is indicated with dashed lines in the lower plot. RIGHT: PARAFAC2 scores and elution time loadings. Note that fifteen loading profiles are obtained per factor calculated by the PARAFAC2 model. Mass spectral loadings are not shown, but these were found similar between the two models.

The data shown in Figure 32 illustrate that for slightly shifted data PARAFAC is able to find reasonable elution time loadings that resembles the loadings from the model of non-shifted data. This would have been even worse for more severe shifts. The compensation for the shifted profiles can be seen in the negative contributions in the loadings. Even though reasonable elution time and correct mass spectral loadings are obtained (the latter not shown), the score plot shows that samples are poorly separated into the known classes compared to Figure 27. For PARAFAC2 the situation is different. PARAFAC2 provides an efficient description of all the shifted profiles. With correct mass spectral profiles obtained at the same time, the resulting score plot provides the proper clustering of the mixture samples. The multiple loading profiles obtained in PARAFAC2 increases the number of elements that must be estimated and thus the degrees of freedom used by the model. This is exemplified further in Table 5.

Table 5. Number of elements estimated in the modes using different chemometric models. This issue has also been addressed in Smilde and Doornbos (1991) for unfold PCA and PARAFAC.

Model	Data	Sample mode [1]	Elution time mode	Mass spectral mode
PCA (one sample landscape) – Figure 24	$J \times K$ [2]		Loadings ($J \times R$)	Scores ($K \times R$)
PCA (summed MS) – Figure 25 and 26	$I \times J$	Scores ($I \times R$)	Loadings ($J \times R$)	
Unfold PCA (not illustrated)	$I \times JK$	Scores ($I \times R$)	Loadings ($J \times K \times R$)	
PARAFAC – Figure 27 and 28	$I \times J \times K$	Scores ($I \times R$)	Loadings ($J \times R$)	Loadings ($K \times R$)
PARAFAC2 – Figure 30 and 32	$I \times J \times K$	Scores ($I \times R$)	Loadings ($I \times J \times R$)	Loadings ($K \times R$)

[1] The parameters calculated for the sample mode are referred to as scores to make the comparison between models easier.
[2] The index of the first dimension is J to refer to a one sample system with elution time and mass spectra indices.

A hypothetical example of a three-way array of 15×51×100 using two components/factors (R = 2) provides models with the following number of elements:

PCA (one sample landscape): 302
PCA (summed MS): 132
Unfold PCA: 10230
PARAFAC1: 332
PARAFAC2: 1760

This indicates that even though PARAFAC2 is a more complex model with more elements to estimate, the model provides similar scores (sample mode) and loadings (mass spectral mode) plots with respect to number of elements as in PARAFAC. The increased number of elements in the elution time mode provides a better description of the real chromatographic data than can be found using any of the strict linear (bi/tri) models. This gives the chemometricians a visual and easy interpretable tool to convince the chromatographers that

PARAFAC2, although a more complex model, can be used to describe real chromatography even in situation with co-eluted and severely shifted peaks.

5.2.4 Determining the proper number of factors/components

Determining the proper number of factors is a crucial step when using multiway models. As mentioned earlier, the PARAFAC model will provide a unique solution, but only when the proper number of factors is determined the unique solution will be chemically meaningful.

For two of the examples presented in Figure 33 (A and B) there is little doubt of how many factors to include in the PARAFAC models, but for the third example (C) it is difficult to see whether two peaks are present or if a single peak is shifted along the elution time axis.

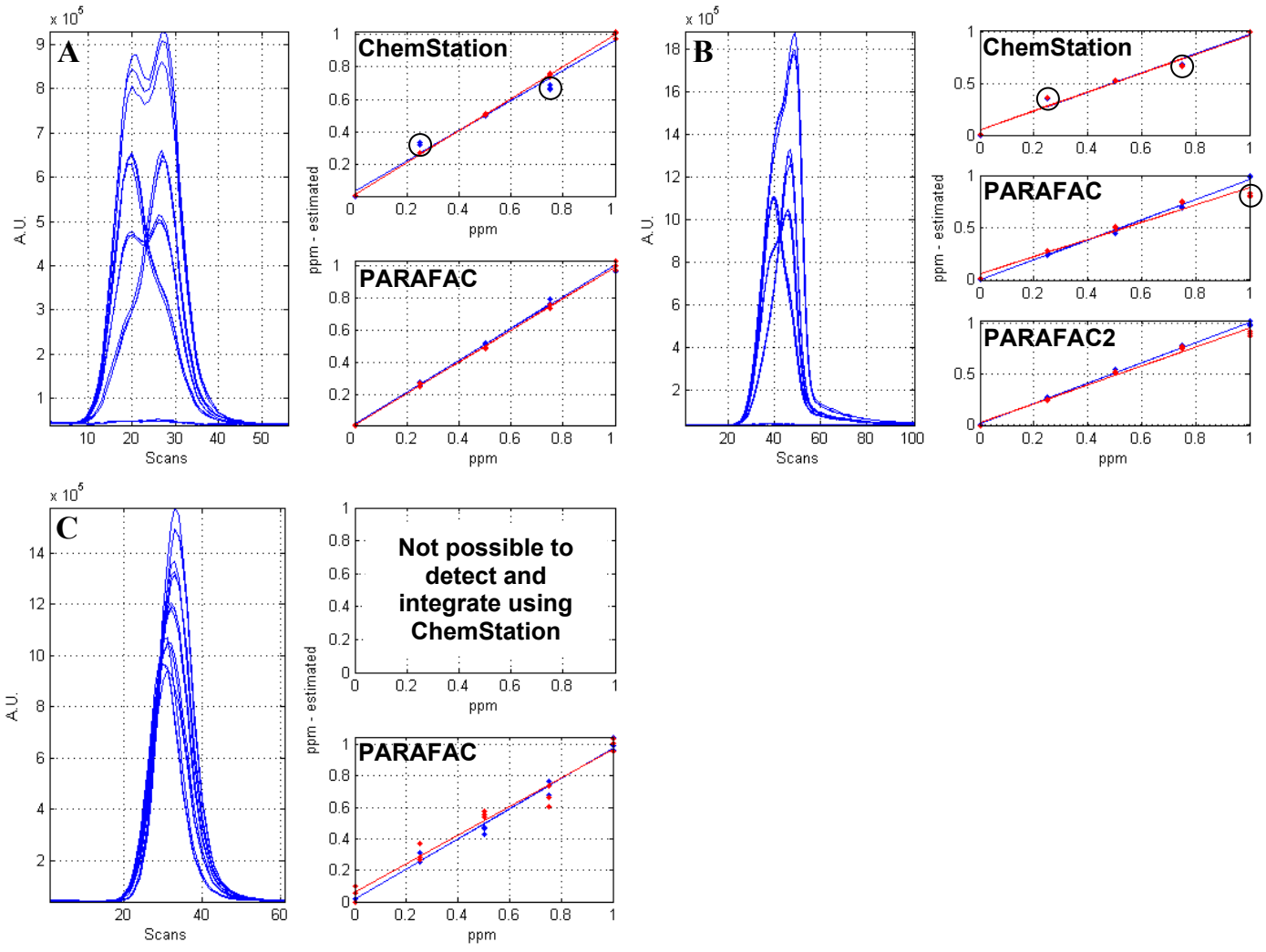


Figure 33. Calibration models for the estimated peak areas using Integration (ChemStation) or score values (PARAFAC/PARAFAC2). In A-C the left plot shows the TIC chromatograms of the three-way array for the peak analyzed. A: See Figure 1 for details. Top: ChemStation, Bottom: PARAFAC. B: 3-methyl butanol and 2-methyl butanol: Top: ChemStation, Middle: PARAFAC, Bottom: PARAFAC2 and C: Diacetyl and 2-pentanone. Top: ChemStation, Bottom: PARAFAC. The circles highlight the differences between the two approaches as explained in the text. Figure from PAPER II.

Extracting only one factor in example (C) would not describe the two known analytes properly. Thus, it is important to be able to determine the complexity of the system and how many factors to calculate. Several approaches can be used for multiway models. As presented earlier, an initial PCA or Evolving Factor Analysis (EFA) on a single sample can help to discover different elution time profiles in different mass channels and from this the rank of the system.

More dedicated methods for multiway data arrays have also been suggested. As for any multivariate model, the model performance can be investigated by adding an additional factor and evaluate the model diagnostics such as the residuals. One of the more recent methods works by calculating a so-called core consistency [Bro and Kiers, 2003]. This diagnostic tool takes into consideration a core array (of dimensions $R \times R \times R$, where R is the number of factors in the model) and it can be proven (Bro and Kiers, 2003) that having a perfect PARAFAC model the superdiagonal of this core array consists of ones with off-superdiagonal elements of zero. If overfitting occurs more off-superdiagonal elements will be non-zero and this lowers the core consistency that is close to 100% for a perfect PARAFAC model. This core array resembles the core array used in the Tucker3 model. If off-superdiagonal non-zero elements are present, this indicates that interactions between different factors from different modes are important (e.g. in a two-factor model described variance could be related to the interaction between the first score vector in **A**, the first loading vector in **B** and the second loading vector in **C**). These interactions violate the trilinear assumptions of the PARAFAC model.

An earlier technique suggested by Harshman and Lundy (1984) is called split-half analysis and uses the intrinsic properties of PARAFAC of finding common parallel proportional profiles in several samples. By splitting the three-way array in two halves (e.g. in the sample mode) and calculating two independent PARAFAC models, two sets of similar loadings are obtained if the proper number of factors is selected. When overfitting occurs these two sets of loadings will be different.

A third more dedicated method was presented by Hoggard and Synovec, (2007). This method evaluates the so-called degenerate solution that can be observed for PARAFAC models with too many factors. A typical sign of a degeneracy is that two of the components become almost identical but with opposite sign or contribution to the model [Bro, 1998]. This was used calculating a so called match value (similarity measure) between a known mass spectrum and the extracted mass spectral loading. A high value indicates that the right analyte is being described by the specific component, but having two high values in the same PARAFAC model indicates a degenerate solution i.e. overfitting. This method is further described and applied in PAPER V.

5.2.5 Advanced methods vs. commercial chromatographic software

For resolved and baseline separated peaks, commercial chromatographic software packages are highly favored due to their easy of operation, point and click control and straightforward identification of analytes based on a library search. However, commercial chromatographic software has many drawbacks compared to advanced chemometric methods:

1) For severely shifted peaks, the software will have problems finding and matching peaks (across samples) falling outside of a user predefined window.
2) If peaks have abnormal shapes the peak start and stop points can be difficult to determine, which causes incorrect baseline estimation when connecting these two points by a straight line (often the default baseline correction method).
3) If peaks elute on top of a gradient or abnormal background behavior the baseline correction approach mentioned above will again be insufficient.
4) With overlapping peaks with different peak height ratio (same height, shoulder peak, tailing peak, absent peak etc.) peak integration can give areas that are underestimated or overestimated (Figure 33 and PAPER II).
5) Often the integration must be validated manually and even for similar samples/products verification of the integration of peaks must be done quite frequently.
6) Factor models or advanced deconvolution techniques are only rarely implemented in software, although AMDIS (Automated Mass spectral Deconvolution and Identification System) is an exception. AMDIS offers a more advanced approach of controlling and monitoring overlapping peaks but relies on selective ions [Dromey et al., 1976; Stein, 1999].
7) The visualization of unique elution time and mass spectral profile separated from baseline contributions are inadequate and often only the extracted peak area can be saved and exported.

All these aspects have been addressed so far in this thesis and using advanced mathematical methods can solve these problems rather efficiently. For the chromatographers, the methods might still seem rather complex, but understanding that the methods simply correct for instrumental artifacts, that they model and present real chemistry and that all this can be visually controlled should be convincing enough.

In PAPER II it was shown that PARAFAC and PARAFAC2 were superior for normal peak integration using default settings in the ChemStation GC software [Agilent, 2001]. In Figure 33 calibrations models from the same mixture system mentioned previously and two other co-elution peaks regions (same samples) have been shown. The findings confirm the difference between ordinary peak integration (Figure 10 in Chapter 2) and using factor models for overlapping peaks of different heights. Ordinary peak integration suffers from the improper peak division that causes an overestimation or underestimation of the peak areas.

These problems are handled easily with PARAFAC and PARAFAC2 and in the case with changes in peak shape (Figure 33B) PARAFAC2 showed better potential as expected.

5.3 Advantages and thoughts when using advanced multivariate models

Several advantages of using multivariate techniques have been put forward so far and to sum up these include (with the multivariate technique mentioned in brackets); explorative/what happens in my data and why (PCA), quantification and identification through peak deconvolution (PARAFAC, PARAFAC2), chemical interpretation of loadings (PARAFAC, PARAFAC2), in-model shift handling (PARAFAC2), fingerprinting (PCA) and visualization of chromatography (all multivariate methods, but especially multiway methods for chemically meaningful chromatography). To explore multidimensional data, the so-called Tucker3 model could also be applied. This multiway model allows interactions between components of different modes and extract features using orthogonality constraints (like PCA) [Henrion, 1994; Tucker, 1966]. Thus, no unique solutions are achieved with Tucker3, but interesting interactions can often be found for data structures that do not comply with the PARAFAC model. The use of Tucker3 for chromatographic data will not be further studied here, but examples can be found in (Cocchi et al., 2008; de Juan and Tauler, 2001; Garcia et al., 2004; Pravdova et al., 2002a). The most straightforward advantage of using multivariate techniques not mentioned so far is the *noise reduction* obtained as a consequence of using more (redundant) measurements of the same phenomenon [Bro, 2003; Lee et al., 1991]. As each chromatogram will contain some sample dependent noise the first extracted loading will be a weighted average of the commonality between chromatograms (e.g. a specific peak shape) and the scores would be the amount of this weighted average. The individual noise not pertaining to the common profile will then be found in the residual matrix. From this it can be seen that the common extracted profile (latent variable) only describes what is in common and will be less noisy than looking at individual raw chromatograms. This also suggests that taking more sample chromatograms into consideration gives a more robust estimate of this latent variable and an increased noise reduction. This means that the signal-to-noise ratios of the modeled peaks are increased compared to classic peak integration using the raw more noisy data [PAPER IV].

In PAPER IV, PARAFAC2 was tested for analysis of different types of chromatographic peak systems from GC-MS measurements of wine samples [PAPER III]. Here PARAFAC2 successfully resolved severely overlapping peaks and modeled peaks of low intensity. However, it was also found that when PARAFAC2 is applied to a system with a large co-eluted peak and a peak of low intensity then some additional consideration must be done to deconvolute all three peaks. The first step was to separate the two peak systems (co-eluted peak and low-intense peak). This gave superior results for the peak of low intensity. The co-eluted peak was possible but more difficult to resolve due to the similar mass spectral profile

of the two analytes. The results from PAPER IV are mentioned here as they raise important issues using PARAFAC or PARAFAC2 for peak modeling.

1) What is a low-rank system?
2) How does PARAFAC model shoulders and embedded peaks?
3) How different can peak shapes be when using PARAFAC2?
4) To what extent can very low-intense peaks be modeled – PARAFAC vs. PARAFAC2?
5) Baseline correction as an in-model step?

The answers to these questions depend on many things (e.g. data structure, number of samples, signal to-noise- ratios, rank etc.) and inherently no explicit answers can be given. The intention with this part is simply to highlight areas related to data and models that need some consideration (e.g. by the chromatographer) before or when the advanced methods are used.

What is a low-rank system?

All factor models operate by describing the main part of the systematic variance (information) in a data array by a reduced set of extracted loadings. As mentioned, these loadings can be derived from a peak system to provide either mathematical and/or chemical solutions. This reduced set of extracted loadings describes a low-rank system and the complexity of this is established when the proper number of components used have been determined. For PARAFAC, a low-rank system is directly translated as a solution of rank-one contributions (individual contributions of analyte-unique loadings of elution time and mass spectral profile) from each analyte found in the peak system. But, there is a limit for how many analytes that can be modeled in each PARAFAC model if meaningful results should still be obtained.

A low rank peak system for GC-MS data is normally a system containing as a maximum a handful of analytes and often only a single peak is being modeled depending on the efficiency in the peak finding algorithm [Arroyo et al., 2007; Comas et al., 2004; Ebrahimi et al., 2007; Johnson et al., 2004]. Throughout this PhD study, several PARAFAC and PARAFAC2 models have been calculated from different data structures and complexities. The general findings have indicated that the more complex nature of the chromatographic data both with respect to co-eluted peaks and to difference between mass spectra, the simpler the information in the peak systems must be in order to get meaningful results. The latter means that more focus must be put on finding peaks and setting the boundaries for the peak systems to be modeled. For GC-MS experiments using EI ionization several fragments in the obtained mass spectra would be similar between analytes and a large number of analytes (either co-eluted or resolved) would be problematic. In a study by Bylund et al. (2002) with LC-MS using electrospray ionization (ESI), seven analytes were easily indentified and

modeled by PARAFAC. The ESI ion source is a soft ionization method providing few fragments that result in simple mass spectra with high intensity of the molecular ion.

How does PARAFAC model shoulder and embedded peaks?

A shoulder peak can be difficult to integrate if ordinary integrations methods like commercial GC software with classical peak separation methods are used (e.g. drop, tangent – Figure 10 in Chapter 2). In Figure 33A and Figure 34, examples of this have been shown. Here both a superior calibration model and proper estimation of the unique elution time profile are achieved. The mass spectral loadings also show superior correlation to the analytes (confirmed by library search).

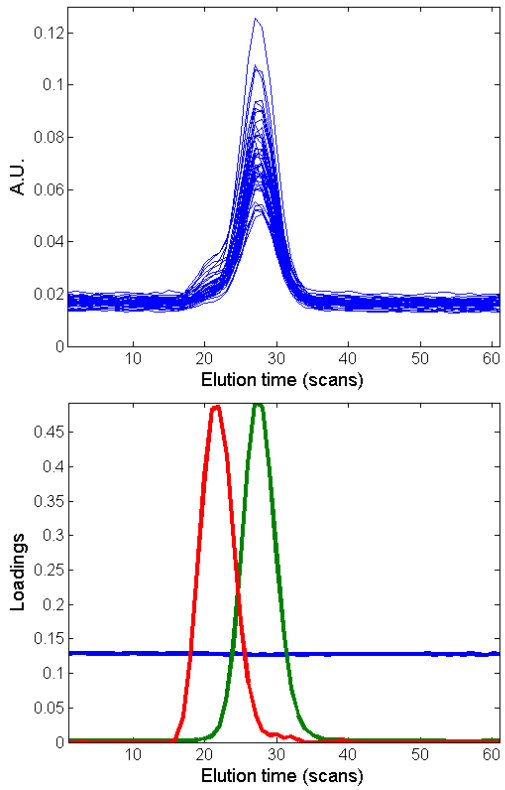


Figure 34. Illustration of fitting a PARAFAC models to a peak system consisting of two overlapping peaks with different peak heights. Here the small peak is just visible as a shoulder on the high-intense peak. TOP: Raw data with two analytes and a baseline offset (TIC chromatogram). BOTTOM: Elution time loadings for the three extracted contributions.

If peaks are totally merged (e.g. a small peak totally embedded in a large broader peak) then just looking at the TIC chromatogram might not reveal the presence of the embedded peak. Looking at specific fragments/ions could reveal this if selective ions are present in the mass spectra. In this case more advanced software (e.g. AMDIS) could be applied and knowledge of the embedded peak might be found from individual mass channels. However, PARAFAC

handles this efficiently and in Figure 35 such an embedded peak is shown together with parameters from the applied PARAFAC model.

For other factor models (e.g. MCR) these situations can be difficult to fit, but including additional samples with a less severe overlap (small part of one peak is present outside the other) would provide a possibility to find unique contributions [de Juan and Tauler, 2007; Manne, 1995]. In GC-MS totally embedded peaks are rather uncommon and as such the problem of embedded peaks is rare whereas shoulder peaks often complicates the normal peak integration procedure.

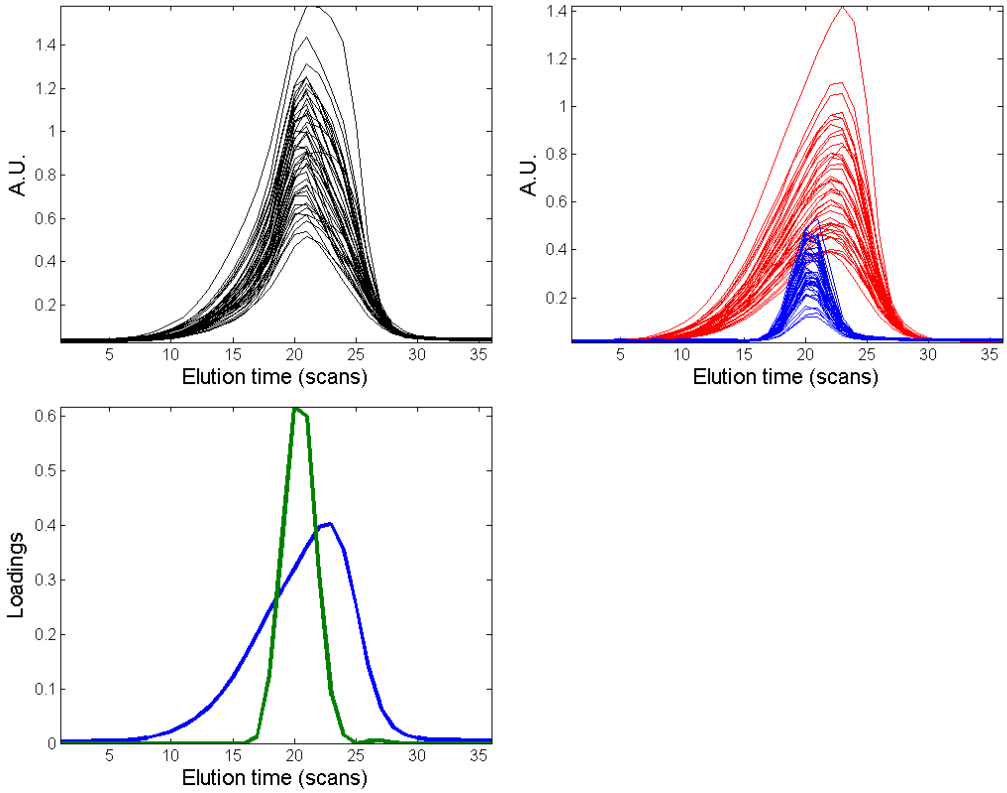


Figure 35. Illustration of fitting a PARAFAC model to a two-analyte peak system where one peak is embedded in the other peak (simulated data). The two analytes in the peak system (blue and red curves) have different mass spectra. TOP – left: Raw data (sum of the signal of two analytes) (TIC chromatograms), right: Raw data split into the two contributions showing the embedded smaller peaks within the larger peak (TIC chromatograms). BOTTOM: The elution time loadings using non-negativity constraints in the chromatographic and mass spectral mode.

How different can peak shapes be when using PARAFAC2?

As explained earlier for PARAFAC2, the inner product of the shifted elution time profiles $\mathbf{B}_i$, individual $\mathbf{B}_i$ for each sample, should be constant as $\mathbf{B}_i^T\mathbf{B}_i = \ldots = \mathbf{B}_I^T\mathbf{B}_I$, for $i = 1,\ldots,I$. As shown earlier shifted peak profiles along the elution time axis are allowed and in the example similar peak features were found in the shifted profiles (Table 4 and Figure 31). But, these peak features can also change to some extent and still be modeled efficiently by

PARAFAC2. This is an important property of PARAFAC2 as a peak from the same analyte can have a different peak shape across samples. E.g. if very high and low concentrated samples are modeled some peak tailing can be observed for the most concentrated samples whereas the less concentrated samples provide more Gaussian shaped peaks. In Figure 36 an example of this has been shown. In this case severe peak shape changes across samples for the same analyte show that even when peak shapes change drastically from sample to sample the PARAFAC2 model finds unique elution time profiles for all samples ($\mathbf{B}_i^T\mathbf{B}_i$ was also confirmed to be the similar across samples (for $i = 1,\ldots I$) – results not shown).

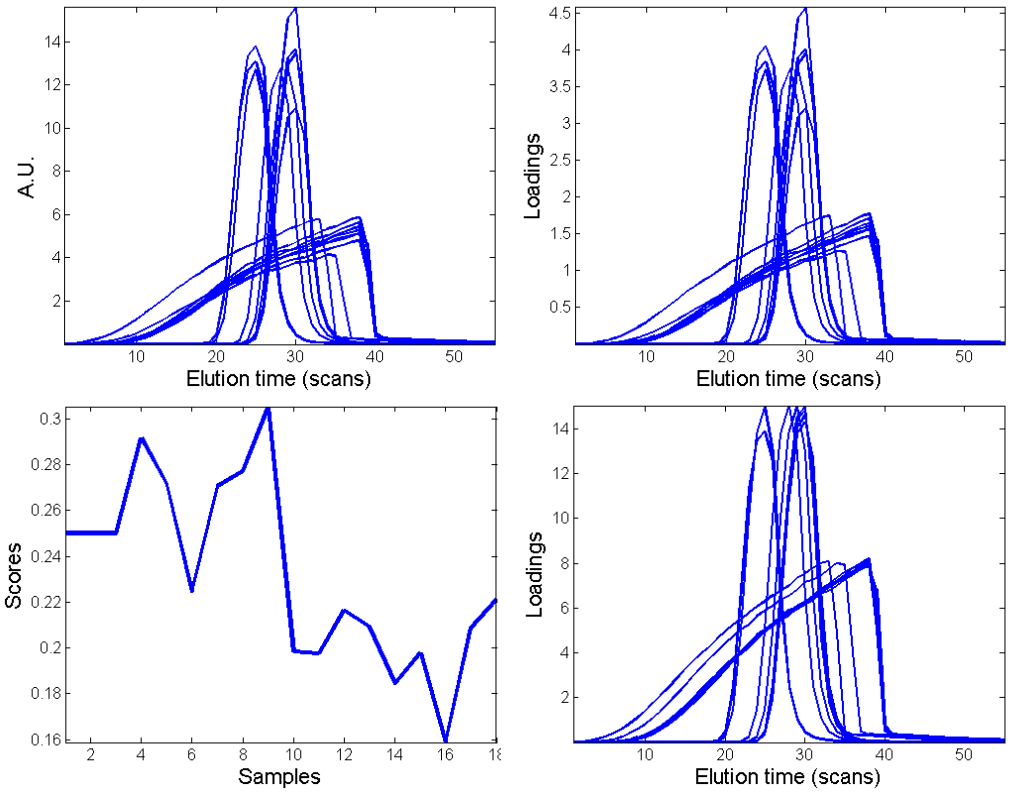


Figure 36. Visualization peak shape changes of one peak of the same analyte for several samples. TOP – left: Raw GC-MS data (TIC chromatograms) holding one peak from the same analyte (same mass spectrum) for several samples. Right: Elution time loadings multiplied with its corresponding score value from a one-factor PARAFAC2 model. BOTTOM – left: Scores and right: Elution time loadings.

The peak shape changes shown in Figure 36 are more severe than would be expected in real chromatographic measurements. For chromatographic data, a 'rule of thumb' for a peak system (whether it be a rank-one system or a more complex but still low-rank system) is that if the baseline is reached before and after the peak then PARAFAC2 is able to model the system even with severe peak shape changes. In these situations the peak information is started and ended in the peak region and this seems to stabilize the solution. Naturally the degree of noise, the number and similarity of analytes in the peak system (complexity) and the changes in peak shape will affect the capability of PARAFAC2 to model the peak system adequately. For kinetic data and batch data, where both unfinished profiles, profiles of

different length and of different shapes can be obtained, the challenges are more significant, but also here PARAFAC2 has shown to be successful [Cueva et al., 2001; Wise et al., 2001].

To what extent can very low-intense peaks be modeled – PARAFAC vs. PARAFAC2?

For this section and the following some highlights from PAPER IV and PAPER V are presented to discuss the issues raised earlier. For GC-MS data, Amigo et al. (PAPER IV) showed that peaks of very low signal-to-noise ratio could be modeled by PARAFAC2 in a two-analyte peak system. For single analyte systems, PARAFAC2 performs well and as shown in PAPER V, PARAFAC2 is not as sensitive as PARAFAC and more affected by noise. This was also expected as PARAFAC2 uses more degrees of freedom especially in the mode where individual profiles are modeled (see also Table 5). For GC×GC-TOFMS data, Skov et al. (PAPER V) found that PARAFAC2 was capable of both identifying and estimating the correct relative concentration for analytes in a known concentration of 1×10^{-8} g/mL. PARAFAC was superior and found a correct relative concentration when the analyte was analyzed in a concentration of 3×10^{-9} g/mL. The correctness of the estimated relative concentrations was determined from how much they deviated from the 'true' calibration line estimated using higher concentrations. The concentrations mentioned here were the lowest amounts of analyte providing an estimate of the relative concentration not deviating significantly from the calibration line.

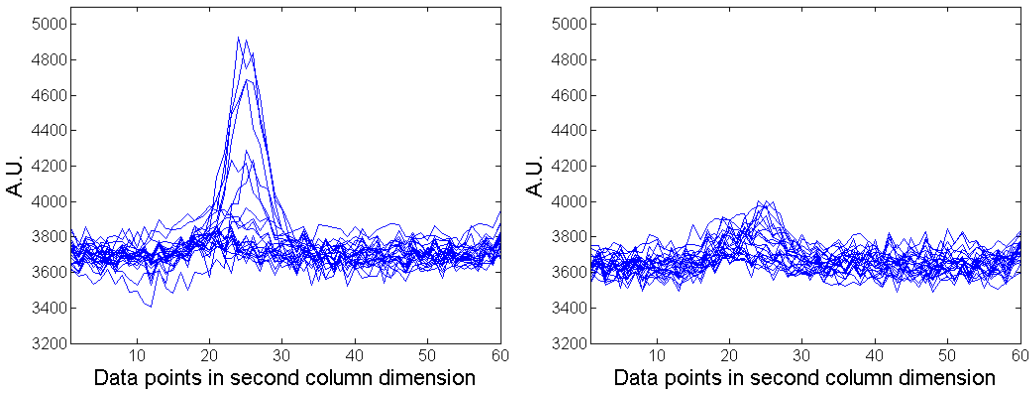


Figure 37. Illustration of a local peak (bromobenzene) from a GC×GC-TOFMS measurement in LEFT: 1×10^{-8} g/mL and RIGHT: 3×10^{-9} g/mL. See PAPER V for further information about instrument settings and concentration range of bromobenzene.

For the lower concentration of bromobenzene in Figure 37, PARAFAC2 was capable of finding and quantifying the analyte with a slight deviation from the calibration line [PAPER V]. These findings confirm that PARAFAC is indeed more sensitive and as such if data is not shifted PARAFAC outperforms PARAFAC2.

Baseline correction as an ‘in-model’ step?

In Chapter 4 different baseline correction methods were presented and it was mentioned that for multidimensional chromatographic data the baseline correction may be handled within the model step. An example of this was shown in Figure 29 for PARAFAC and an additional example is presented in Figure 38 for PARFAC2. In Figure 38, two PARAFAC2 components describe the contributions from the two analytes and one component holds the baseline contribution and the remaining (the noise) are then gathered in the residuals. Using multiway models for baseline correction, the baseline contributions can change between samples as long as the mass spectrum is rather constant over the entire elution time region included. The latter is often observed from benchtop GC-MS instruments where a reproducible Electron Ionization (EI) source is used.

For the PARAFAC2 model the unexplained part holds the noise separated from the analytical and the baseline signal. The effect of baseline correction (and de-noising) can be evaluated by reconstructing the analytical signal using only the contributions from the two analytes extracted from the low-rank peak system.

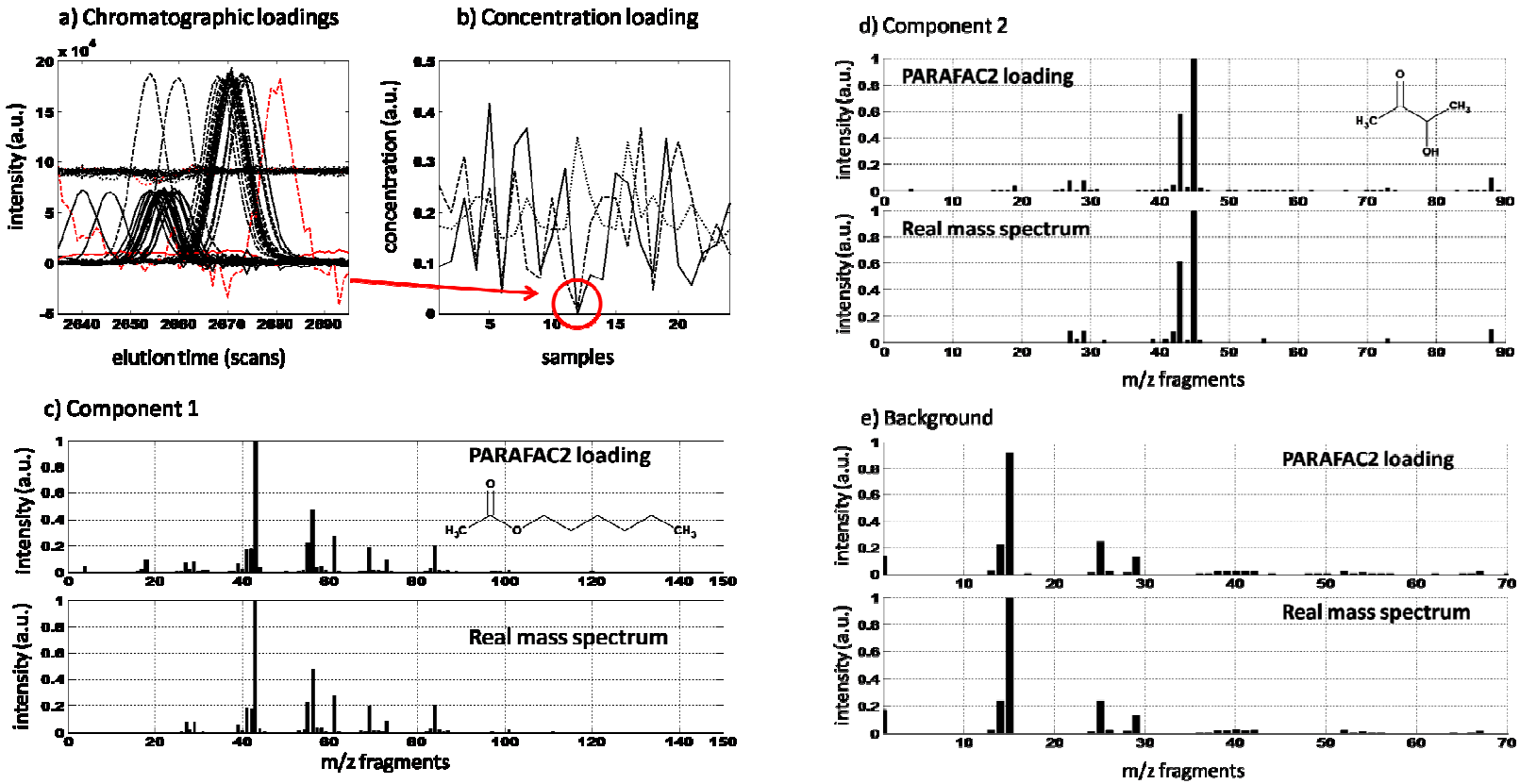


Figure 38. Example of PARAFAC2 'baseline correction' and 'de-noising'. a) Chromatographic loadings (three factor model) and b) concentration loading (background, dotted; factor 1, solid; factor 2, dashed). Comparison between PARAFAC2 mass spectral loading (upper) and real mass spectrum (bottom) for c) factor 1: acetic acid hexyl ester, d) factor 2: 3-hydroxy-2-butanone and e) factor 3: the background. The real mass spectra were found from running standards of analytes after they had been identified in ChemStation [Agilent, 2001]. Figure is modified from PAPER IV.

6 Information in multiple data blocks

After proper pre-processing and deconvolution, aligned chromatographic profiles and/or peak areas are readily available. These multiple data blocks can hold different parts of information for a specific purpose. Some parts can be redundant, some unique and some not even descriptive for the given purpose.

To evoke the title of this thesis; mathematical resolution of complex chromatographic measurements, this refers to all aspects of data pretreatment and modeling. This was all covered in PAPER I, II, IV and V and have been mentioned in the chapters presented so far. The last step is then to use this readily available information and look for patterns, characteristics and deviations among the samples. This naturally leads the way for a method to compare information in pre-processed data available from GC-MS or other types of measurements.

This chapter contains aspects of evaluating multiple data sets containing different parts of information. By analyzing a set of samples using multiple chromatographic techniques e.g. GC-MS for flavor descriptions and LC-MS for taste descriptions, much information is obtained. Part of this information may be unique to an individual data block or found in both data blocks. For GC-MS data, the TIC chromatograms, the mass spectral profiles and the peak areas are three data blocks that can be obtained from GC-MS experiments as described earlier. The peak areas could be found from the deconvolution methods presented in Chapter 5. E.g. by combing a peak finding routine and the advanced multivariate methods (PAPER II) or by using commercial GC software [Agilent, 2001].

Comparing or extracting information from multiple data blocks is termed multiblock analysis. Several multiblock methods have been described in literature and have shown to be successful when comparing information in different data blocks [Berglund and Wold, 1999; Felicio et al., 2005; Hoskuldsson and Svinning, 2006; PAPER III; Tenenhaus and Vinzi, 2005; Westerhuis et al., 1998; Westerhuis and Smilde, 2001; Wold et al., 1996]. In general, multiblock methods focus on optimizing the predictive power of extracted features of the combined multiple blocks. This could be with respect to finding the best prediction of the class relationship of the wines from the two different geographical regions mentioned earlier.

In PAPER III a technique denoted Multiblock Variance Partitioning (MVP), to objectively compare information found in different data blocks, is presented. The more traditional multiblock approach will not be explained further here. The method described in PAPER III works by finding unique, common and unexplained parts comparing multiple data blocks as

illustrated in Figure 39 (see PAPER III for further details about the mathematical equations behind MVP).

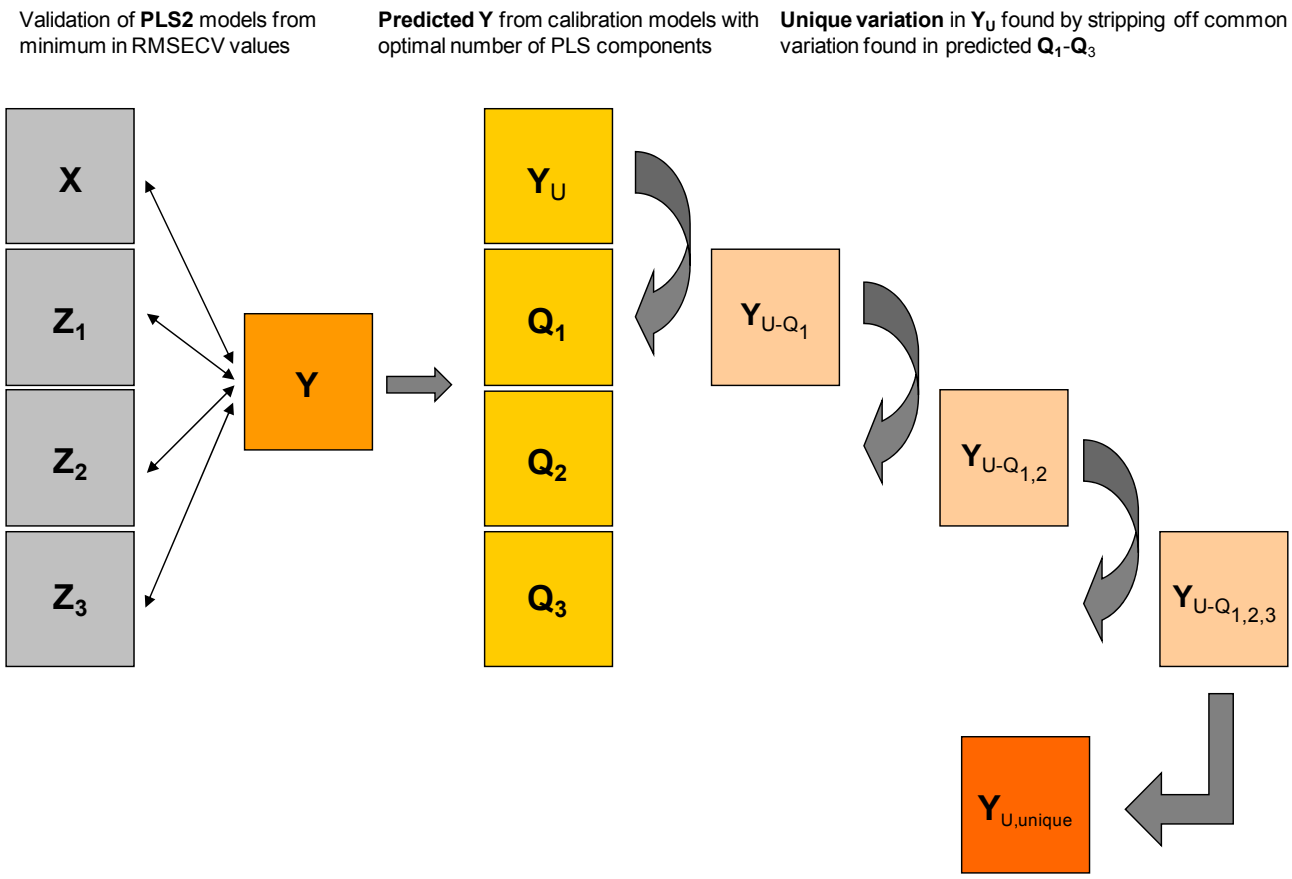


Figure 39. Diagram illustrating the Multiblock Variance Partitioning – MVP approach. The sequential removal of common variation is shown to make the visualization more clear. The correct mathematical equations are given in section 2.5., where further details of MVP can also be found. $Y_{U,unique}$ is the same as $Y_{U\text{-}Q_{1,2,3}}$. From PAPER III.

The MVP approach is based on PLS regression models between the block to be predicted (**Y**) and the blocks that predict (**X**/**Z**) with **X** being the base block. The base block is the data block for which we will find parts of information that is either only found in this block (unique) or is common with one or more blocks (**Z**). PLS will not be explained further here but information can be found in (Esbensen, 2002; Geladi and Kowalski, 1986; Naes et al., 2002).

To show the effect of MVP for GC-MS data, the wine data set presented and used in PAPER III will be pre-processed and modeled using a selection of the methods presented so far (data block 1) and compared to three other data blocks; block 2) peak areas integrated by ChemStation, block 3) the elution time profiles (TIC chromatograms) and block 4) the mass spectral profiles (see PAPER III for details about the latter three data blocks).

All four blocks depend on treating data in or exporting them from ChemStation (here used as the commercial chromatographic software).

Block 2: In ChemStation peaks are identified based on their mass spectra (or selected ions/fragments) and located in each sample from local searches within narrow windows along the elution time dimension. Severely shifted peaks can hamper this peak locating routine and force the windows to be inappropriately wide. At worst, this can result in similar peaks being wrongly assigned. Thus, it would be of great interest to be able to align data beforehand. So far this is not possible for GC-MS data, but has been solved for vectorized signals like GC-FID using e.g. LineUp [Infometrix, 2006]. Here data can be exported to LineUp and aligned and then imported back to ChemStation. Another major limitation is the peak division tools available. These were addressed in Figure 10. Often the simple drop method is applied but then one has to accept the known errors when integrating overlapping peaks (Figure 33).

The other three blocks depend on data being transferred to a numeral software program such as MATLAB for enhanced data manipulation. Exporting data from ChemStation is not a trivial task. Nevertheless, this can be handled with proper toolboxes and one way of doing it has been put forward in PAPER II.

For **block 3** and **4** one dimension is removed from the GC-MS data by collapsing either the mass spectral or the elution time dimension (shown in
Figure **11**).

One way of achieving **block 1** is to use a handful of the pre-processing and modeling methods presented so far.

1) Alignment
    a. Only small shifts were observed and this can be handled easily with any of the techniques mentioned. Here COW was used on the TIC chromatograms and for GC-MS the found warping path used for individual mass channels.
2) Baseline correction
    a. For the TIC chromatogram the baseline correction using splines (van den Berg, 2008) was applied with the settings mentioned in Figure 21.
    b. For landscapes (GC-MS) no baseline correction was applied as this was assumed to be handled in the multiway models, as explained earlier.
3) Peak finding in the TIC chromatogram
    a. A simple peak finding method based on peak intensities above a certain threshold was used and peak regions were estimated from derivatives as explained by Vivo-Truyols et al. (2005). In this way several peak systems

positioned along the elution time dimension holding one or more peaks are established.

4) Modeling of peak systems
    a. Identification and quantification of individual analytes using PARAFAC (or PARAFAC2 depending on the success of the alignment method).
    b. The number of components was validated using core-consistency and estimated elution time loadings. Also EFA could be applied here for an initial guess of the rank of the peak system.
5) Scores from models used as relative concentrations.
6) All scores are divided with the score of the internal standard added (normal procedure for GC peak integration studies).
7) The data (from now on denoted the '*advanced'* block) can now be found as a table with wines as objects and aroma compounds as variables.
8) The information in the four blocks can be compared using MVP.

These data are rather homogenous and simple, meaning that all samples contain the same analytes and that the number of peaks is moderate with baseline regions between the majorities of the peaks. This of course makes the alignment, baseline correction and peak finding more straightforward and make the selection of input parameters less crucial. This also makes it possible to automate the many steps/routines so that similar data can be handled easily. The automation issue should always be addressed when several individual steps are a prerequisite for subsequent data exploration. As shown chromatographic data can be of significantly different complexity and although similar pre-processing steps can be used, the steps require different methods and settings. To come up with a universal toolbox is probably overly ambitious, but rather dedicated subroutines can be made for specific types of data structures and complexities.

For each local peak region a PARAFAC model is run with up to three factors. As co-elution of three peaks is not expected, three factors are satisfactory to model two overlapping peaks plus a likely baseline contribution. Two examples of this semi-automatic peak modeling approach have been shown in Figure 40.

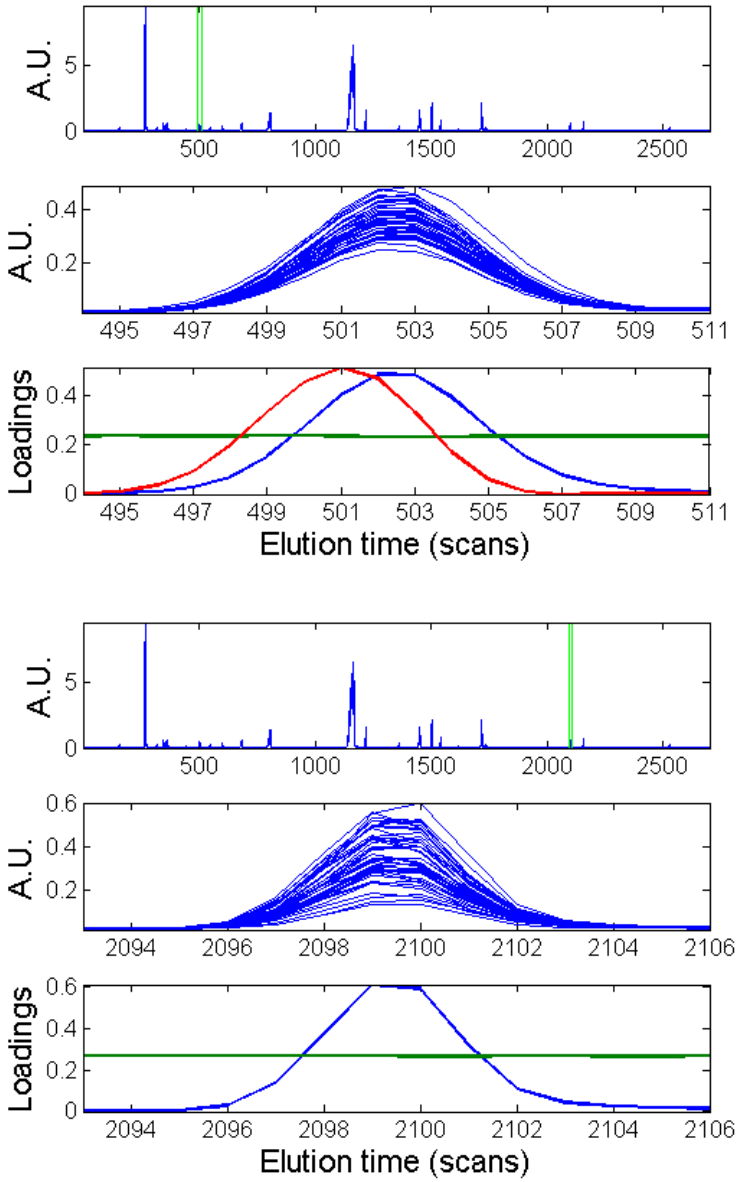


Figure 40. Local peak regions analyzed using a PARAFAC model with one to three factors as explained in the text. TOP: A peak system of two analytes difficult to see from the shown TIC chromatogram (and also when zooming in – middle plot). The three loadings for the two analytes and baseline contributions are shown. BOTTOM: Similar as top, but now a two-factor model was found to be optimal based on the core-consistency (dropped when using three factors) value and the elution time loadings.

The information in the four data blocks are compared using the MVP approach presented in detail in PAPER III. All four data blocks are used as both **X** and **Y** blocks and the results are shown in Figure 41.

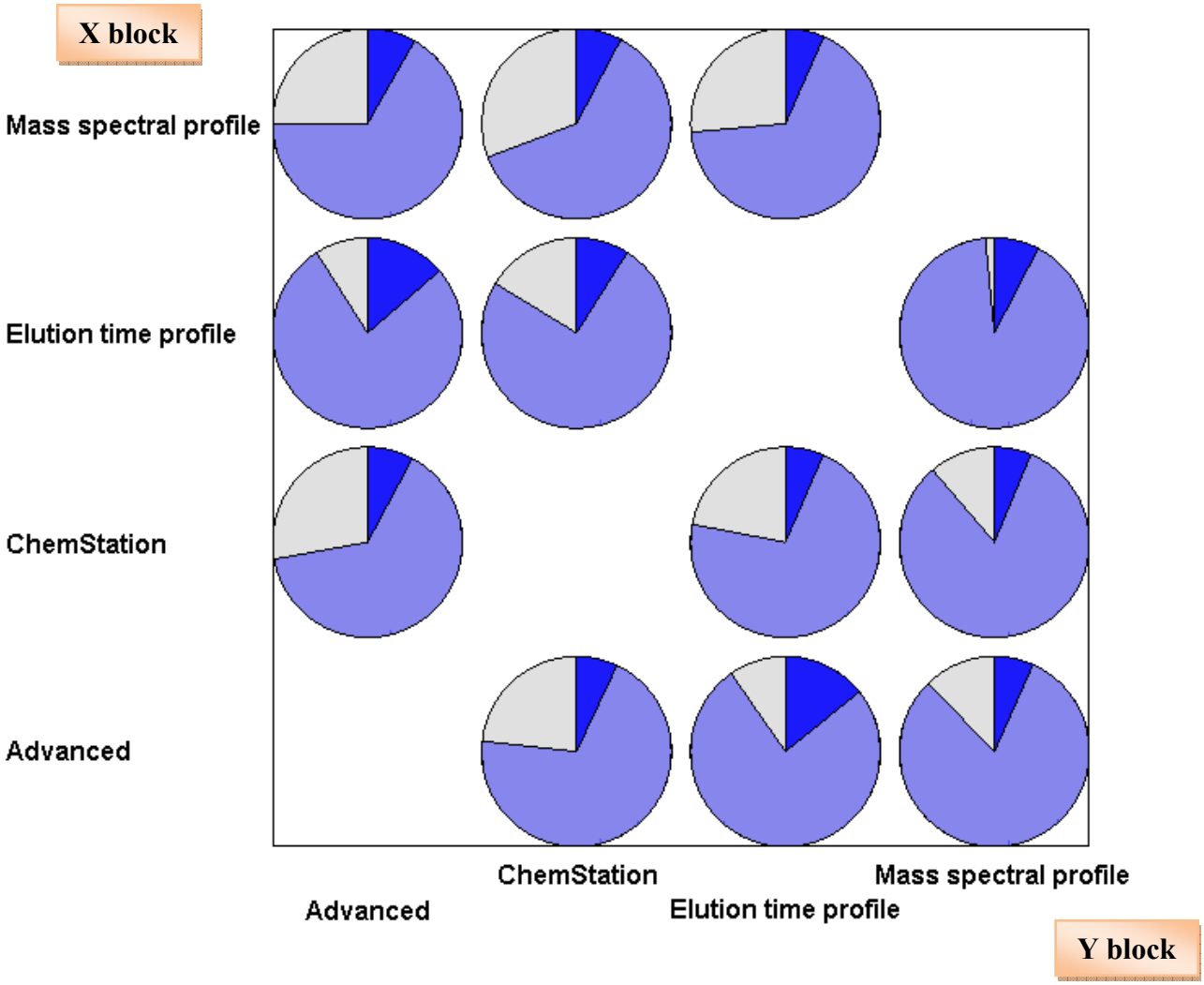


Figure 41. MVP approach used for four data blocks: Mass spectral and elution time profiles from raw data and peak areas estimated from commercial GC software (ChemStation with default conditions) and using advanced methods presented in this PhD thesis. Interpretation of MVP: Unexplained part (light grey), common part (light blue) and unique part (dark blue) in percentages of the variation in **Y**. See PAPER III for a detailed description of the circles and their parts.

This small example captures some rather interesting things. First of all, the four data blocks contain information only related to the specific block. This part is rather small, which was also expected as the four data blocks are from the same initial GC-MS data. Secondly, although all data blocks are derived from the same data source, the explained variance of **Y** is ranging from 70% to almost 100% and a higher percent than 70% could have been expected. However, this might be explained by the additional noisy variables found in especially Elution time and Mass spectral profiles and different peaks integrated using ChemStation and Advanced method. Also, for ChemStation some uncertainty in the peak integration of overlapping peaks can be expected, as shown earlier (Figure 33).

Three data blocks describe the individual peaks; Advanced and ChemStation hold the integrated peak areas and Elution time profiles hold a global fingerprint with peak intensities. The mass spectral profiles are different, as all information about the peaks has been mixed up. Thus, not surprisingly the Mass spectral profiles are found to the block that in general describes the other data blocks most inefficiently (top row in Figure 41).

Finally, the peak areas from the advanced methods are equal to or better at predicting the elution time and mass spectral profiles (fingerprints) compared to commercial GC software; ChemStation (higher explained variance). Thus, it seems as more (relevant) variance is kept in the advanced block descriptive for the ChemStation block than vice versa. More examples and interpretations of MVP results have been discussed in PAPER III.

7 Conclusions and perspectives

In the process of implementing new chemometric methods and approaches to solve specific addressed problems, scientists from different disciplines are brought together. I feel that this interdisciplinary nature of CHEMOMETRICS is its inner strength. It acts as a catalyst to creativity and problem-solving, and by simply bringing together motivated and skilled persons the product of their mutual efforts will be much more than the sum of the parts[10].

This thesis has mainly focused on developing and applying advanced data analysis tools for gas chromatographic measurements. Two main subjects have been investigated: 1) pre-processing of chromatographic data and 2) modeling of chromatographic data. It has been the intention to strengthen the link between chromatography and the available mathematical tools (chemometrics) in order to exploit the many possibilities that are presently available. Perhaps the most central conclusion is that chromatographers and chemometricians can benefit greatly from each other. This collaboration is essential for developing dedicated techniques, which may lead to implementation of advanced chemometric methods in commercial chromatographic software.

A more detailed account on specific results and some of the perspectives on topics treated are given in the ensuing sections.

7.1 Pre-processing of chromatographic data

The fact that pre-processing is important for chromatographic data is well-known and have been demonstrated throughout this thesis. For chromatographic data, instrumentally induced artifacts like background and shifted peaks have been studied and several suggestions for how to handle these have been presented. These instrumental artifacts complicate the extraction of the relevant analytical signal by introducing additional contributions to the signal, which can be difficult to distinguish from minor chemical compounds.

For alignment techniques a handful of the most frequently used methods for chromatographic data have been presented and described. Among these methods, COW is by far the most investigated and has proven successful for alignment of both simple and complex chromatograms. The high interest in COW has spread its use to commercial software packages as well (e.g. LineUp – Infometrix, 2006). The primary uses and modifications of COW are still located in the chromatographic research environments, but the use is spread more and more to other scientific fields than chromatography where shifted profiles are obtained (e.g. NMR, sensory science and studies over time).

[10]Epilogue found in Brakstad (1995).

The popularity of COW was one of the motivations that initiated the development of the automated alignment method presented in PAPER I. In PAPER I a simplex coordinate optimization routine was presented that optimized novel quantitative measures for alignment quality and peak area preservation. These measures were calculated from the best aligned data using different input parameters (segment length and slack size) in the COW algorithm. In this way a near-optimal alignment was found in significantly less time than needed for a trial and error approach. A similar optimization routine may be applied for the PWA methods presented where similar input parameters are needed. One prerequisite is that the global search space shows a clear and smooth pattern for the two quantitative measures. For COW it was demonstrated that large segments and low flexibility resulted in poor alignment but a preserved peak area and that small segments and large flexibility gave a good alignment but also peak shape changes. In between these extremes a clear and smooth pattern, although with local deviations, was observed. Although, the PWA methods have been found to be capable of correcting less complex shifts the speed of the algorithms makes the use of multiple (and optimized) alignment runs even more appealing.

In the alignment methods presented, little attention has been given to making sure that the intensity over all time points are kept intact. One method for doing this was presented in PAPER I using a combined measure of the sample-wise deviation in Euclidean length for sample chromatograms before and after alignment. However, this measure is calculated for the whole sample set and changes in e.g. different samples for important local peak regions can be missed if at the same time major peaks are preserved. New ideas for how to handle this should be pursued to ensure that peak areas can be even better preserved. This is of course most important when peak areas are the final objective.

7.2 Modeling of chromatographic data

Multivariate data analysis methods (PCA, PARAFAC and PARAFAC2) have been demonstrated on chromatographic data of different dimensionality and structure. The multiway models have been demonstrated to be advantageous for GC-MS and GC×GC-TOFMS data exploiting the second order nature of the data. Which model to use depends on the properties of the data and which kind of information should be extracted. When using PARAFAC and PARAFAC2 for peak systems containing single, resolved or overlapped peaks, chemically meaningful profiles were extracted. The chromatographic and mass spectral profiles where found to be true estimates of the individual contributions of the analytes present in the peak system. For more difficult peaks systems with severe co-elution, a smaller peak observed as a shoulder or embedded peak was efficiently modeled with PARAFAC.

Despite the success for shoulder and embedded peak, selectivity is still an issue using these models. Having totally overlapped peaks of similar shapes or almost identical mass spectra

the individual contributions will be nearly impossible to extract without some kind of data manipulations. Also if peak intensity ratios (e.g. between two co-eluted peaks) are constant across samples this will be difficult to model, but can be solved by including an additional sample where this peak ratio is different. However, as shown even embedded peaks can be handled by PARAFAC, as long as a distinct pattern in peak shape and mass spectrum are found between the embedded and the surrounding peak.

The determination of the proper number of factors was demonstrated to be a crucial step in multivariate and especially multiway modeling. Several methods were put forward to evaluate and estimate the complexity of a certain peak system; methods that could also be included in an automated modeling routine. Being able to find the peaks along the elution time axis and determine the complexity of local peak systems, peak areas in the properly pre-processed chromatograms were readily accessible from individual PARAFAC or PARAFAC2 models. Such an approach was presented for rather homogeneous and simple chromatographic data. Combined with mathematical validation measures and chromatographic knowledge (e.g. visual confirmation of elution time and mass spectral profiles) the peaks areas were obtained in a semi-automated way.

7.3 Information in multiple data blocks

The multiblock method (MVP) presented in PAPER III and described in Chapter 6 objectively compares variance found in multiple data blocks. The method was shown to provide an overview of differences and commonalities between data blocks and this could be illustrated in a straightforward and simple manner. Compared to other multiblock methods, MVP does not bring information related to individual or combinations of variables responsible for the description of the **Y** block. This must be explored after the MVP calculation and as such no improvement of the prediction error can be achieved. Being able to elucidate what kind of information the unique part holds for different **X** blocks will provide an increased understanding of the relationship between the data blocks investigated. If the information in one **Z** block is intrinsically the same as in the **X** block, then the unique part of **X** when predicting **Y** will be close to zero. This can make the interpretation rather difficult. However, this is rarely a big problem if the experiments have been designed properly (e.g. NIR instruments placed in strategic positions in the process) or analytical techniques known to focus on different properties of the samples are used.

7.4 Outro

Taking into consideration the many advantages of applying multiway analysis for chromatographic data, one may wonder why these powerful methods are still generally underused in the daily work and how this can be solved. The answer to this was partly addressed earlier with the major issue being how this could be automated and more user-

friendly. It would be overambitious to be able to handle all problems observed for chromatographic data by one universal toolbox. Rather, dedicated toolboxes for specific problems and data types seem to be the right way to go.

Research should be done to automate as much as possible of this advanced methodology and to provide user-friendly software that does not require much decision making.

Inclusion of these advanced methods in commercial instrumentation software would make the techniques more familiar to a broader range of end-users. But before this can be a reality, feedback from chromatographers or other experts in chromatographic measurements is needed. This will link chromatography and chemometrics in the best possible way.

By developing and discussing ways to automate some of the existing techniques used for chromatographic data handling, this thesis is a step in the direction of more user-friendly software. One of the next steps should be convincing chromatographers that chemometricians can help solving many of their daily data problems. Introducing the mathematical methods as a complementary, but equally important, method to the instrumental settings would enhance the overall chromatographic performance significantly.

8 References

Adahchour, M., Beens,J., Vreuls,R.J.J., and Brinkman,U.A.T. **2006a**. Recent developments in comprehensive two-dimensional gas chromatography (GC X GC) I. Introduction and instrumental set-up. *Trac-Trends in Analytical Chemistry*, 25 (5): 438-454.

Adahchour, M., Beens,J., Vreuls,R.J.J., and Brinkman,U.A.T. **2006b**. Recent developments in comprehensive two-dimensional gas chromatography (GC x GC) II. Modulation and detection. *Trac-Trends in Analytical Chemistry*, 25 (6): 540-553.

Agilent. ChemStation - GCD Plus. [A.01.00]. **2001**. Palo Alto, CA, USA, Hewlett Packard.

Arroyo, D., Ortiz,M.C., and Sarabia,L.A. **2007**. Multiresponse optimization and parallel factor analysis, useful tools in the determination of estrogens by gas chromatography-mass spectrometry. *Journal of Chromatography A*, 1157 (1-2): 358-368.

Bassompierre, M., Tomasi,G., Munck,L., Bro,R., and Engelsen,S.B. **2007**. Dioxin screening in fish product by pattern recognition of biomarkers. *Chemosphere*, 67 (9): S28-S35.

Berglund, A. and Wold,S. **1999**. A serial extension of multiblock PLS. *Journal of Chemometrics*, 13 (3-4): 461-471.

Bicking, M.K.L. **2006a**. Integration errors in chromatographic analysis, part I: Peaks of approximately equal size. *Lc Gc North America*, 24 (4): 402-414.

Bicking, M.K.L. **2006b**. Integration errors in chromatographic analysis, part II: Large peak size ratios. *Lc Gc North America*, 24 (6): 604-616.

Bobleter, O. **1996**. Exhibition of the first gas chromatographic work of Erika Cremer and Fritz Prior. *Chromatographia*, 43 (7-8): 444-446.

Booksh, K.S. and Kowalski,B.R. **1994**. Theory of Analytical-Chemistry. *Analytical Chemistry*, 66 (15): A782-A791.

Boque, R. and Ferre,J. **2004**. Using second-order data in chromatographic analysts. *Lc Gc Europe*, 17 (7): 402-407.

Brakstad, F. **1995**. The Feasibility of Latent-Variables Applied to Gc-Ms Data. *Chemometrics and Intelligent Laboratory Systems*, 29 (2): 157-176.

Brereton, R.G. **1995**. Deconvolution of Mixtures by Factor-Analysis. *Analyst*, 120 (9): 2313-2336.

Bro, R. **1997**. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38 (2): 149-171.

Bro, R. **1998.** Multi-Way Analysis in the Food Industry. University of Amsterdam, Thesis/Dissertation.

Bro, R. **2003**. Multivariate calibration - What is in chemometrics for the analytical chemist? *Analytica Chimica Acta*, 500 (1-2): 185-194.

Bro, R. **2006**. Review on multiway analysis in chemistry - 2000-2005. *Critical Reviews in Analytical Chemistry*, 36 (3-4): 279-293.

Bro, R., Andersson,C.A., and Kiers,H.A.L. **1999**. PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *Journal of Chemometrics*, 13 (3-4): 295-309.

Bro, R. and Kiers,H.A.L. **2003**. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17 (5): 274-286.

Bro, R. and Smilde,A.K. **2003**. Centering and scaling in component analysis. *Journal of Chemometrics*, 17 (1): 16-33.

Bro, R., Smilde,A.K., and de Jong,S. **2001**. On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 58 (1): 3-13.

Bylund, D., Danielsson,R., Malmquist,G., and Markides,K.E. **2002**. 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 (2): 237-244.

Carroll, J.D. and Chang,J.J. **1970**. Analysis of Individual Differences in Multidimensional Scaling Via An N-Way Generalization of Eckart-Young Decomposition. *Psychometrika*, 35 (3): 283-&.

Christensen, J.H., Hansen,A.B., Karlson,U., Mortensen,J., and Andersen,O. **2005a**. Multivariate statistical methods for evaluating biodegradation of mineral oil. *Journal of Chromatography A*, 1090 (1-2): 133-145.

Christensen, J.H. and Tomasi,G. **2007**. Practical aspects of chemometrics for oil spill fingerprinting. *Journal of Chromatography A*, 1169 (1-2): 1-22.

Christensen, J.H., Tomasi,G., and Hansen,A.B. **2005b**. Chemical fingerprinting of petroleum biomarkers using time warping and PCA. *Environmental Science & Technology*, 39 (1): 255-260.

Cocchi, M., Durante,C., Grandi,M., Manzini,D., and Marchetti,A. **2008**. Three-way principal component analysis of the volatile fraction by HS-SPME/GC of aceto balsamico tradizionale of modena. *Talanta*, 74 (4): 547-554.

Comas, E., Gimeno,R.A., Ferre,J., Marce,R.M., Borrull,F., and Rius,F.X. **2004**. Quantification from highly drifted and overlapped chromatographic peaks using second-order calibration methods. *Journal of Chromatography A*, 1035 (2): 195-202.

Cueva, J.M.D., Rossi,A.V., and Poppi,R.J. **2001**. Modeling kinetic spectrophotometric data of aminophenol isomers by PARAFAC2. *Chemometrics and Intelligent Laboratory Systems*, 55 (1-2): 125-132.

Daszykowski, M. and Walczak,B. **2006**. Use and abuse of chemometrics in chromatography. *Trac-Trends in Analytical Chemistry*, 25 (11): 1081-1096.

Daszykowski, M. and Walczak,B. **2007**. Target selection for alignment of chromatographic signals obtained using monochannel detectors. *Journal of Chromatography A*, 1176 (1-2): 1-11.

de Juan, A. and Tauler,R. **2001**. Comparison of three-way resolution methods for non-trilinear chemical data sets. *Journal of Chemometrics*, 15 (10): 749-772.

de Juan, A. and Tauler,R. **2007**. Factor analysis of hyphenated chromatographic data - Exploration, resolution and quantification of multicomponent systems. *Journal of Chromatography A*, 1158 (1-2): 184-195.

Debeljak, Z., Srecnik,G., Madic,T., Petrovic,M., Knezevic,N., and Medic-Saric,M. **2005**. Evaluation of novel sample identification approach based on chromatographic fingerprint set correlation homogeneity analysis. *Journal of Chromatography A*, 1062 (1): 79-86.

Dixon, S.J., Brereton,R.G., Soini,H.A., Novotny,M.V., and Penn,D.J. **2006**. An automated method for peak detection and matching in large gas chromatography-mass spectrometry data sets. *Journal of Chemometrics*, 20 (8-10): 325-340.

Dolan, J.W. **2002**. Peak tailing and resolution. *Lc Gc Europe*, 15 (6): 334-337.

Dolan, J.W. **2003**. Why do peaks tail? *Lc Gc Europe*, 16 (9): 610-615.

Dromey, R.G., Stefik,M.J., Rindfleisch,T.C., and Duffield,A.M. **1976**. Extraction of Mass-Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography Mass Spectrometry Data. *Analytical Chemistry*, 48 (9): 1368-1375.

Duarte, A.C. and Capelo,S. **2006**. Application of chemometrics in separation science. *Journal of Liquid Chromatography & Related Technologies*, 29 (7-8): 1143-1176.

Ebrahimi, D., Li,J.F., and Hibbert,D.B. **2007**. Classification of weathered petroleum oils by multi-way analysis of gas chromatography-mass spectrometry data using PARAFAC2 parallel factor analysis. *Journal of Chromatography A*, 1166 (1-2): 163-170.

Eilers, P.H.C. **2004**. Parametric time warping. *Analytical Chemistry*, 76 (2): 404-411.

Esbensen, K. 2002. *Multivariate Data Analysis - In Practice*. Camo Process AS. Olso, Norway.

Escandar, G.M., Faber,N.K.M., Goicoechea,H.C., de la Pena,A.M., Olivieri,A.C., and Poppi,R.J. **2007**. Second- and third-order multivariate calibration: Data, algorithms and applications. *Trac-Trends in Analytical Chemistry*, 26 (7): 752-765.

Ettre, L.S. **1993**. Nomenclature for Chromatography. *Pure and Applied Chemistry*, 65 (4): 819-872.

Ettre, L.S. **2000**. Chromatography: the separation technique of the 20th century. *Chromatographia*, 51 (1-2): 7-17.

Ettre, L.S. **2003**. M.S. Tswett and the invention of chromatography. *Lc Gc Europe*, 16 (9): 632-640.

Ettre, L.S. **2008**. The beginnings of gas adsorption chromatography 60 years ago. *Lc Gc North America*, 26 (1): 48-60.

Felicio, C.C., Bras,L.P., Lopes,J.A., Cabrita,L., and Menezes,J.C. **2005**. Comparison of PLS algorithms in gasoline and monitoring with MIR and NIR. *Chemometrics and Intelligent Laboratory Systems*, 78 (1-2): 74-80.

Fellinger A. **1998a**. "Noise." In Attila Fellinger, editor, *Data analysis and signal processing in chromatography*. Elsevier. Amsterdam. 125-141.

Fellinger A. **1998b**. "Peak detection." In Attila Fellinger, editor, *Data analysis and signal processing in chromatography*. Elsevier. Amsterdam. 183-190.

Fellinger A. **1998c**. "Signal enhancement." In Attila Fellinger, editor, *Data analysis and signal processing in chromatography*. Elsevier. Amsterdam. 143-181.

Foley, J.P. **1987**. Systematic-Errors in the Measurement of Peak Area and Peak Height for Overlapping Peaks. *Journal of Chromatography*, 384: 301-313.

Forshed, J., Schuppe-Koistinen,I., and Jacobsson,S.P. **2003**. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487 (2): 189-199.

Gan, F., Ruan,G.H., and Mo,J.Y. **2006**. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82 (1-2): 59-65.

Garcia, I., Sarabia,L., Ortiz,M.C., and Aldama,J.M. **2004**. Three-way models and detection capability of a gas chromatography-mass spectrometry method for the determination of clenbuterol in several biological matrices: the 2002/657/EC European Decision. *Analytica Chimica Acta*, 515 (1): 55-63.

Geladi, P. and Kowalski,B.R. **1986**. Partial Least-Squares Regression - A Tutorial. *Analytica Chimica Acta*, 185: 1-17.

Gong, F., Liang,Y.Z., Fung,Y.S., and Chau,F.T. **2004**. Correction of retention time shifts for chromatographic fingerprints of herbal medicines. *Journal of Chromatography A*, 1029 (1-2): 173-183.

Gorecki, T., Harynuk,J., and Panic,O. **2004**. The evolution of comprehensive two-dimensional gas chromatography (GC x GC). *Journal of Separation Science*, 27 (5-6): 359-379.

Grob, R.L. and Barry,E.F. 2006. *Modern Practice of Gas Chromatography*. John Wiley and Sons Inc. New Jersey, US.

Gross, J.H. 2006. *Mass spectrometry: A Textbook*. Springer. Berlin Heidelberg, New York, US.

Harshman, R.A. **1972**. PARAFAC2 - Mathematical and Technical Notes. *UCLA Working Papers in Phonetics*, 22: 30-44.

Harshman, R.A. **1970**. Foundations of the PARAFAC procedure: Model and conditions for an explanatory multi-mode factor analysis. *UCLA working papers in phonetics*, 16: 1-84.

Harshman, R.A. and Lundy,M.E. **1984**. "The PARAFAC model for three-way factor analysis and multidimensional scaling." In H.G.Law, C.W.Snyder, J.A.Hattie, and R.P.McDonald, editors, *Research methods for Multimode data analysis*. New York, US. Praeger.

Hendriks, M.M.W.B., Cruz-Juarez,L., Bont,D.D., and Hall,R. **2005**. Preprocessing and exploratory analysis of chromatographic profiles of plant extracts. *Analytica Chimica Acta*, 545 (1): 53-64.

Henrion, R. **1994**. N-Way Principal Component Analysis - Theory, Algorithms and Applications. *Chemometrics and Intelligent Laboratory Systems*, 25 (1): 1-23.

Hinshaw, J.V. **2001**. When Good Columns Go Bad. *Lc Gc North America*, 19 (6): 596-603.

Hinshaw, J.V. **2004**. Finding a needle in a haystack. *Lc Gc North America*, 22 (10): 990-997.

Hoggard, J.C. and Synovec,R.E. **2007**. Parallel Factor Analysis (PARAFAC) of Target Analytes in GCxGC-TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors. *Analytical Chemistry*, 79 (4): 1611-1619.

Hoskuldsson, A. and Svinning,K. **2006**. Modelling of multi-block data. *Journal of Chemometrics*, 20 (8-10): 376-385.

Hotelling, H. **1933**. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24: 417-441.

Infometrix. LineUp. [Version 2.03]. **2006**. Infometrix, Inc., Bothell, WA.

Itakura, F. **1975**. Minimum Prediction Residual Principle Applied to Speech Recognition. *Ieee Transactions on Acoustics Speech and Signal Processing*, AS23 (1): 67-72.

James, A.T. and Martin,A.J.P. **1952**. Gas-Liquid Partition Chromatography - the Separation and Micro-Estimaton of Volatile Fatty Acids from Formic Acid to Dodecanoic Acid. *Biochemical Journal*, 50 (5): 679-690.

Johnson, K.J., Rose-Pehrsson,S.L., and Morris,R.E. **2004**. Monitoring diesel fuel degradation by gas chromatography-mass Spectroscopy and chemometric analysis. *Energy & Fuels*, 18 (3): 844-850.

Jonsson, P., Gullberg,J., Nordstrom,A., Kusano,M., Kowalczyk,M., Sjostrom,M., and Moritz,T. **2004**. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry*, 76 (6): 1738-1745.

Kiers, H.A.L. **2000**. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14 (3): 105-122.

Kiers, H.A.L., Ten Berge,J.M.F., and Bro,R. **1999**. PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics*, 13 (3-4): 275-294.

Krebs, M.D., Tingley,R.D., Zeskind,J.E., Holmboe,M.E., Kang,J.M., and Davis,C.E. **2006**. Alignment of gas chromatography-mass spectrometry data by landmark selection from complex chemical mixtures. *Chemometrics and Intelligent Laboratory Systems*, 81 (1): 74-81.

Lee, G.C. and Woodruff,D.L. **2004**. Beam search for peak alignment of NMR signals. *Analytica Chimica Acta*, 513 (2): 413-416.

Lee, T.A., Headley,L.M., and Hardy,J.K. **1991**. Noise-Reduction of Gas-Chromatography Mass-Spectrometry Data Using Principal Component Analysis. *Analytical Chemistry*, 63 (4): 357-360.

Liang, Y.Z., Kvalheim,O.M., Rahmani,A., and Brereton,R.G. **1993**. A Two-Way Procedure for Background Correction of Chromatographic Spectroscopic Data by Congruence Analysis and Least-Squares Fit of the Zero-Component Regions - Comparison with Double-Centering. *Chemometrics and Intelligent Laboratory Systems*, 18 (3): 265-279.

Liang, Y.Z., Wu,H.L., Shen,G.L., Jiang,J.H., Liang,S., and Yu,R.Q. **2006**. Aspects of recent developments in analytical chemometrics. *Science in China Series B-Chemistry*, 49 (3): 193-203.

Liu, Z.Y. and Phillips,J.B. **1991**. Comprehensive 2-Dimensional Gas-Chromatography Using An On-Column Thermal Modulator Interface. *Journal of Chromatographic Science*, 29 (6): 227-231.

Macnaughtan, D., Rogers,L.B., and Wernimon,G. **1972**. Principal-Component Analysis Applied to Chromatographic Data. *Analytical Chemistry*, 44 (8): 1421-1427.

Maeder, M. **1987**. Evolving Factor-Analysis for the Resolution of Overlapping Chromatographic Peaks. *Analytical Chemistry*, 59 (3): 527-530.

Maeder, M. and Zilian,A. **1988**. Evolving Factor-Analysis, a New Multivariate Technique in Chromatography. *Chemometrics and Intelligent Laboratory Systems*, 3 (3): 205-213.

Malmquist, G. and Danielsson,R. **1994**. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *Journal of Chromatography A*, 687 (1): 71-88.

Malmquist, L.M.V., Olsen,R.R., Hansen,A.B., Andersen,O., and Christensen,J.H. **2007**. Assessment of oil weathering by as chromatography-mass spectrometry, time warping and principal component analysis. *Journal of Chromatography A*, 1164 (1-2): 262-270.

Manne, R. **1995**. On the Resolution Problem in Hyphenated Chromatography. *Chemometrics and Intelligent Laboratory Systems*, 27 (1): 89-94.

Marriott, P. and Shellie,R. **2002**. Principles and applications of comprehensive two-dimensional gas chromatography. *Trac-Trends in Analytical Chemistry*, 21 (9-10): 573-583.

Martin, A.J.P. and Synge,R.L.M. **1941**. A new form of chromatogram employing two liquid phases I. A theory of chromatography 2. Application to the micro-determination of the higher monoamino-acids in proteins. *Biochemical Journal*, 35: 1358-1368.

Naes, T., Isaksson,T., Fearn,T., and Davis,T. 2002. *User Friendly Guide to Multivariate Calibration and Classification*. NIR Publications. Chichester, UK.

Nielsen, N.P.V., Smedsgaard,J., and Frisvad,J.C. **1999**. Full second-order chromatographic/spectrometric data matrices for automated sample identification and component analysis by non-data-reducing image analysis. *Analytical Chemistry*, 71 (3): 727-735.

Nielsen, N.P.V., Carstensen,J.M., and Smedsgaard,J. **1998**. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805 (1-2): 17-35.

Norgaard, L., Saudland,A., Wagner,J., Nielsen,J.P., Munck,L., and Engelsen,S.B. **2000**. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54 (3): 413-419.

Ortiz, M.C. and Sarabia,L. **2007**. Quantitative determination in chromatographic analysis based on n-way calibration strategies. *Journal of Chromatography A*, 1158 (1-2): 94-110.

Ovejero-Lopez, I., Bro,R., and Bredie,W.L.P. **2005**. Univariate and multivariate modelling of flavour release in chewing gum using time-intensity: a comparison of data analytical methods. *Food Quality and Preference*, 16 (4): 327-343.

PAPER I. **2006**. Automated alignment of chromatographic data. *Journal of Chemometrics*, 20 (11-12): 484-497.

PAPER II. **2008**. Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry*, 390 (1): 281-285.

PAPER III. **2008**. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Analytica Chimica Acta*, 615 (1): 18-29.

PAPER IV. **2008**. Solving GC-MS problems with PARAFAC2. *Submitted to TrAC Trends in Analytical Chemistry*.

PAPER V. **2008**. Handling shifts in GC×GC-TOFMS data using Shift Correction and Modeling. *In preparation*.

Pearson, K. **1901**. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (7-12): 559-572.

Peris, M. **1996**. An overview of recent expert system applications in analytical chemistry. *Critical Reviews in Analytical Chemistry*, 26 (4): 219-237.

Phillips, J.B. and Beens,J. **1999**. Comprehensive two-dimensional gas chromatography: a hyphenated method with strong coupling between the two dimensions. *Journal of Chromatography A*, 856 (1-2): 331-347.

Pierce, K.M., Hoggard,J.C., Mohler,R.E., and Synovec,R.E. **2008**. Recent advancements in comprehensive two-dimensional separations with chemometrics. *Journal of Chromatography A*, 1184 (1-2): 341-352.

Pierce, K.M., Hope,J.L., Johnson,K.J., Wright,B.W., and Synovec,R.E. **2005**. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *Journal of Chromatography A*, 1096 (1-2): 101-110.

Pierce, K.M., Wright,B.W., and Synovec,R.E. **2007**. Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm. *Journal of Chromatography A*, 1141 (1): 106-116.

Pravdova, V., Boucon,C., de Jong,S., Walczak,B., and Massart,D.L. **2002a**. Three-way principal component analysis applied to food analysis: an example. *Analytica Chimica Acta*, 462 (2): 133-148.

Pravdova, V., Walczak,B., and Massart,D.L. **2002b**. A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, 456 (1): 77-92.

Prince, J.T. and Marcotte,E.M. **2006**. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry*, 78 (17): 6140-6152.

Ray, N.H. **1954a**. Gas Chromatography. I. The Separation and Estimation of Volatile Organic Compounds by Gas-Liquid Partition Chromatography. *Journal of Applied Chemistry*, 4 (1): 21-25.

Ray, N.H. **1954b**. Gas Chromatography. II. The Separation and Analysis of Gas Mixtures by Chromatographic Methods. *Journal of Applied Chemistry*, 4 (2): 82-85.

Reinbach, H.C., Meinert,L., Ballabio,D., Aaslyng,M.D., Bredie,W.L.P., Olsen,K., and Moller,P. **2007**. Interactions between oral burn, meat flavor and texture in chili spiced pork patties evaluated by time-intensity. *Food Quality and Preference*, 18 (6): 909-919.

Rinnan, A., Riu,J., and Bro,R. **2007**. Multi-way prediction in the presence of uncalibrated interferents. *Journal of Chemometrics*, 21 (1-2): 76-86.

Roach, L. and Guilhaus,M. **1992**. Evolving Factor-Analysis in Gas-Chromatography - Mass-Spectrometry - a Feasibility Study. *Organic Mass Spectrometry*, 27 (10): 1071-1076.

Sadygov, R.G., Maroto,F.M., and Huhmer,A.F.R. **2006**. ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Analytical Chemistry*, 78 (24): 8207-8217.

Sakoe, H. and Chiba,S. **1978**. Dynamic-Programming Algorithm Optimization for Spoken Word Recognition. *Ieee Transactions on Acoustics Speech and Signal Processing*, 26 (1): 43-49.

Savitzky, A. and Golay,M.J.E. **1964**. Smoothing + Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36 (8): 1627-1639.

Schmidt, B. **2008.** Chemometric Methods Applied to Herbal Medicinal Products: St. John's Wort. Faculty of Pharmaceutical Sciences, University of Copenhagen, Thesis/Dissertation.

Schmidt, B., Jaroszewski,J.W., Bro,R., Witt,M., and Stark,D. **2008**. Combining PARAFAC Analysis of HPLC-PDA Profiles and Structural Characterization Using HPLC-PDA-SPE-NMR-MS Experiments: Commercial Preparations of St. John's Wort. *Analytical Chemistry*, 80 (6): 1978-1987.

Schoenmakers, P., Marriott,P., and Beens,J. **2003**. Nomenclature and conventions in comprehensive multidimensional chromatography. *Lc Gc Europe*, 16 (6): 335-339.

Sinha, A.E., Hope,J.L., Prazen,B.J., Fraga,C.G., Nilsson,E.J., and Synovec,R.E. **2004a**. Multivariate selectivity as a metric for evaluating comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry subjected to chemometric peak deconvolution. *Journal of Chromatography A*, 1056 (1-2): 145-154.

Sinha, A.E., Prazen,B.J., and Synovec,R.E. **2004b**. Trends in chemometric analysis of comprehensive two-dimensional separations. *Analytical and Bioanalytical Chemistry*, 378 (8): 1948-1951.

Skov, T. and Bro,R. **2005**. A new approach for modelling sensor based data. *Sensors and Actuators B-Chemical*, 106 (2): 719-729.

Smilde, A.K., Bro,R., and Geladi,P. 2004. *Multi-Way Analysis. Applications in the Chemical Sciences*. Wiley. Chichester.

Smilde, A.K. and Doornbos,D.A. **1991**. 3-Way Methods for the Calibration of Chromatographic Systems - Comparing Parafac and 3-Way Pls. *Journal of Chemometrics*, 5 (4): 345-360.

Smolkova-Keulemansova, E. **2000**. A few milestones on the journey of chromatography. *Hrc-Journal of High Resolution Chromatography*, 23 (7-8): 497-501.

Stein, S.E. **1999**. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10 (8): 770-781.

Steinier, J., Termonia,Y., and Deltour,J. **1972**. Comments on smoothing and differentiation of data by simplified least squares procedure. *Analytical Chemistry*, 44 (11): 1906-1909.

Suits, F., Lepre,J., Du,P., Bischoff,R., and Horvatovich,P. **2008**. Two-Dimensional Method for Time Aligning Liquid Chromatography-Mass Spectrometry Data. *Analytical Chemistry*, 80 (9): 3095-3104.

Szymanska, E., Markuszewski,M.J., Capron,X., van Nederkassel,A.M., Heyden,Y.V., Markuszewski,M., Krajka,K., and Kaliszan,R. **2007**. Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides. *Electrophoresis*, 28 (16): 2861-2873.

Tauler, R. **1995**. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems*, 30 (1): 133-146.

Tenenhaus, M. and Vinzi,V.E. **2005**. PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis. *Journal of Chemometrics*, 19 (3): 145-153.

Tomasi, G. **2006.** Practical and computational aspects in chemometric data analysis. Faculty of Life Sciences, University of Copenhagen, http://www.models.life.ku.dk, Thesis/Dissertation.

Tomasi, G., van den Berg,F., and Andersson,C. **2004**. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18 (5): 231-241.

Torgrip, R.J.O., Aberg,M., Karlberg,B., and Jacobsson,S.P. **2003**. Peak alignment using reduced set mapping. *Journal of Chemometrics*, 17 (11): 573-582.

Tswett, M. **1906**. Physikalisch-chemische Studien über das Chlorophyll. Die Adsorptionen. *Ber Deutsch Bot Ges*, 24: 316-323.

Tucker, L.R. **1966**. Some Mathematical Notes on Three-Mode Factor Analysis. *Psychometrika*, 31 (3): 279-311.

van den Berg, F. Baseline_spline: Determines splines-based baseline by gradually eliminating points. **2008**.

van den Berg, F., Tomasi,G., and Viereck,N. **2005**. "Warping: investigation of NMR pre-processing and correction." In S.B.Engelsen, P.S.Belton, and H.J.Jakobsen, editors, *Magnetic Resonance in Food Science: The Multivariate Challenge*. The Royal Society of Chemistry. Cambridge. 131-138.

van Mispelaar, V.G., Tas,A.C., Smilde,A.K., Schoenmakers,P.J., and van Asten,A.C. **2003**. Quantitative analysis of target components by comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, 1019 (1-2): 15-29.

van Nederkassel, A.M., Daszykowski,M., Eilers,P.H.C., and Heyden,Y.V. **2006a**. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118 (2): 199-210.

van Nederkassel, A.M., Daszykowski,M., Massart,D.L., and Vander Heyden,Y. **2005**. Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling. *Journal of Chromatography A*, 1096 (1-2): 177-186.

van Nederkassel, A.M., Xu,C.J., Lancelin,P., Sarraf,M., MacKenzie,D.A., Walton,N.J., Bensaid,F., Lees,M., Martin,G.J., Desmurs,J.R., Massart,D.L., Smeyers-Verbeke,J., and Vander Heyden,Y. **2006b**. Chemometric treatment of vanillin fingerprint chromatograms - Effect of different signal alignments on principal component analysis plots. *Journal of Chromatography A*, 1120 (1-2): 291-298.

Vandenbogaert, M., Li-Thiao-Te,S., Kaltenbach,H.M., Zhang,R., Aittokallio,T., and Schwikowski,B. **2008**. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics.*, 8(4): 650-672.

Vivo-Truyols, G., Torres-Lapasio,J.R., van Nederkassel,A.M., Vander Heyden,Y., and Massart,D.L. **2005**. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals - Part I: Peak detection. *Journal of Chromatography A*, 1096 (1-2): 133-145.

Walczak, B. and Wu,W. **2005**. Fuzzy warping of chromatograms. *Chemometrics and Intelligent Laboratory Systems*, 77 (1-2): 173-180.

Westerhuis, J.A., Kourti,T., and MacGregor,J.F. **1998**. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12 (5): 301-321.

Westerhuis, J.A. and Smilde,A.K. **2001**. Deflation in multiblock PLS. *Journal of Chemometrics*, 15 (5): 485-493.

Willse, A., Belcher,A.M., Preti,G., Wahl,J.H., Thresher,M., Yang,P., Yamazaki,K., and Beauchamp,G.K. **2005**. Identification of major histocompatibility complex-regulated body odorants by statistical analysis of a comparative gas chromatography/mass spectrometry experiment. *Analytical Chemistry*, 77 (8): 2348-2361.

Wilson, J.D. and McInnes,C.A.J. **1965**. The elimination of errors due to baseline drift in the measurement of peak areas in gas chromatography. *Journal of Chromatography A*, 19: 486-494.

Wise, B.M., Gallagher,N.B., Bro,R., Shaver,J.M., Windig,W., and Koch,R.S. PLS Toolbox. [Version 4.1]. **2006**. Eigenvector Research Inc., Manson, WA.

Wise, B.M., Gallagher,N.B., and Martin,E.B. **2001**. Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *Journal of Chemometrics*, 15 (4): 285-298.

Wold, S., Esbensen,K., and Geladi,P. **1987**. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2 (1-3): 37-52.

Wold, S., Kettaneh,N., and Tjessem,K. **1996**. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, 10 (5-6): 463-482.

Wold, S. and Sjostrom,M. **1977**. "SIMCA: A method for analyzing chemical data in terms of similarity and analogy." In B.R.Kowalski, editor, *Chemometrics Theory and Application*. American Chemical Society. Wash., D.C. 243-282.

Wu, W., Daszykowski,M., Walczak,B., Sweatman,B.C., Connor,S.C., Haseldeo,J.N., Crowther,D.J., Gill,R.W., and Lutz,M.W. **2006**. Peak alignment of urine NMR spectra using fuzzy warping. *Journal of Chemical Information and Modeling*, 46 (2): 863-875.

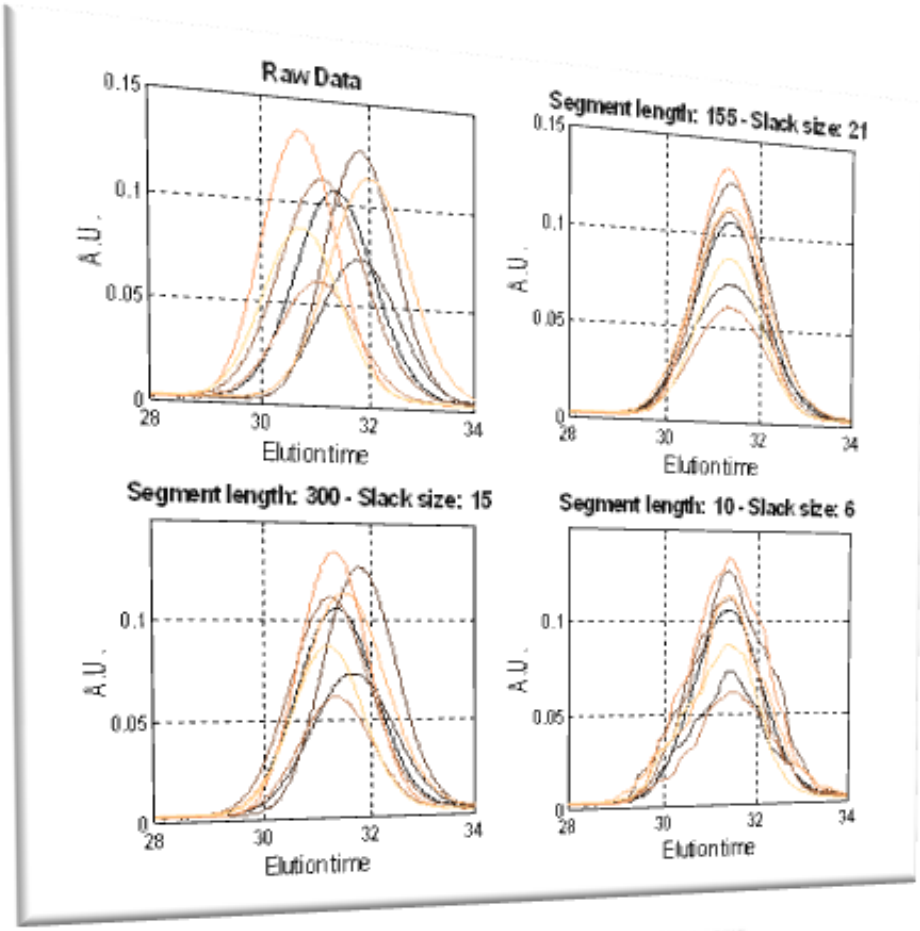
Xu, C.J., Liang,Y.Z., Chau,F.T., and Vander Heyden,Y. **2006**. Pretreatments of chromatographic fingerprints for quality control of herbal medicines. *Journal of Chromatography A*, 1134 (1-2): 253-259.

Yao, W.F., Yin,X.Y., and Hu,Y.Z. **2007**. A new algorithm of piecewise automated beam search for peak alignment of chromatographic fingerprints. *Journal of Chromatography A*, 1160 (1-2): 254-262.

PAPER I-V

PAPER I

Skov, T., van den Berg, F., Tomasi, G., and Bro, R. **2006**. Automated alignment of chromatographic data. *Journal of Chemometrics*, 20 (11-12): 484-497.



JOURNAL OF CHEMOMETRICS
J. Chemometrics 2006; **20**: 484–497
Published online 27 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.1031

Automated alignment of chromatographic data

Thomas Skov*, Frans van den Berg, Giorgio Tomasi and Rasmus Bro

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

Received 20 June 2006; Revised 2 January 2007; Accepted 5 January 2007

This paper focuses on the practical aspects and implications of preprocessing chromatographic data to correct for undesirable time-shifts. An approach to automate the alignment of chromatographic data based on peak alignment or warping is proposed. This approach deals with selection of the required parameters including selection of reference sample to warp towards, and chooses warping settings based on a new evaluation criterion for goodness of correction. The new criterion aims at quantifying goodness of alignment while at the same time penalising significant shape or area-changes in the warped peaks. The entire selection procedure is automated using a discrete-coordinates simplex-like optimisation routine. Examples with simulated chromatographic data, GC-FID and HPLC-Fluorescence measurement series illustrate the potential of using this automated alignment tool. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: automated alignment; correlation optimised warping; peak area preservation; chromatographic data; optimisation

1. INTRODUCTION

Preprocessing of chromatographic data to correct for undesirable phenomena in the signal is often a crucial step in the proper data analysis chain. This holds especially if the data are to be used for multivariate data analysis either in the form of peak areas or raw chromatograms. Some of the 'artifacts' can be taken care of using traditional chromatographic procedures such as correcting the signal by internal standards or normalisation. Other, more challenging artifacts such as peak shifting and baseline variations need more advanced preprocessing techniques to remove their undesired contribution to the subsequent data analysis steps like principal components analysis (PCA) or PARAllel FACtor analysis (PARAFAC). The reason is that if data are not brought to a form where elements in the matrix or data cube for individual objects or samples describe the same phenomena, the required assumption of bi- or tri-linearity in the data is no longer valid. Several preprocessing methods have been put forward in literature to correct for shifted peaks in chromatographic data [1–7].

The correlation optimised warping (COW) algorithm has shown great potential for alignment correction in chromatographic profiles, due to its assumed peak shape and area preserving properties [1–4]. The COW algorithm is based on aligning a sample chromatogram in the form of a digitised vector towards a target chromatogram (i.e. a reference sample vector) by piecewise linear stretching or compression in combination with interpolation, optimising the correlation coefficients between corresponding segments in reference and sample [3]. The same reference sample is used for correcting/aligning the entire data set. As the chromatograms are split in a number of segments and all boundaries between segments are allowed to move a certain number of data points in either direction—the local flexibility of alignment—the COW algorithm requires two user input parameters: the *segment length* and the flexibility (so-called *slack size*). These two parameters are typically selected on a trial and error basis by visual inspection of the chromatographic profiles (peak shape, width, etc.). An automated method to investigate whether selected parameters are optimal has not yet been proposed in the literature. This paper introduces a new concept for this purpose that calculates a so-called *simplicity* value for each combination of input parameters in the COW algorithm [3]. It can be used as a measure of the similarity of the shapes of the aligned chromatograms.

As shown in the Section 3 of this paper, one parameter combination of segment and slack size will give the highest simplicity value, but more combinations will provide simplicity values very close to this optimum. Since an exhaustive search will typically be too time consuming for representative research questions, the latter observation can be utilised in a stratified optimisation procedure that investigates fewer combinations and provides a satisfying

*Correspondence to: T. Skov, Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
E-mail: thsk@life.ku.dk

alignment in considerably less time. However, when aligning peaks consisting of only a relative small number of sample points, the interpolation step can cause a significant change in peak shapes and areas. This paper includes a second optimisation criterion of minimising area alterations in the optimisation routine leading to a more conservative and rational measure of a correct choice of segment length and slack size.

One obstacle in alignment is the selection of the right reference sample, and so far no general method is put forward in the literature that can guide a good choice. The ideal reference sample should be as representative as possible for all phenomena of interest in the data set. This could, for example, be the chromatogram in the middle of the run sequence [2], the chromatogram containing the highest number of common chemical constituents (i.e. peaks) [3–4], a composite artificial sample, or the chromatogram that is most similar to the loading of the first principal component in a PCA model on the un-aligned data set. This paper discusses and demonstrates a simple and quick way of selecting a reference in a given data set based on the product of correlation coefficients.

In this paper, we focus on the practical aspects and implications of preprocessing chromatographic data to correct for undesired time-shifts. Optimisation of the alignment of chromatographic data will be demonstrated on three data sets. A simulated set (Figure 1) will be used to explain the principles of the simplicity value and the importance of preserving the peak area as quantitative measure for an optimal alignment. Next, a GC-FID data set of ground coffee samples (Figure 2) will be used to illustrate the above on real data and to show the effect of selecting a suitable reference vector [3]. Lastly, a HPLC-Fluorescence data set is also included to show the possibilities of the ideas presented here for chromatographic methods with broad-peak features (Figure 3).

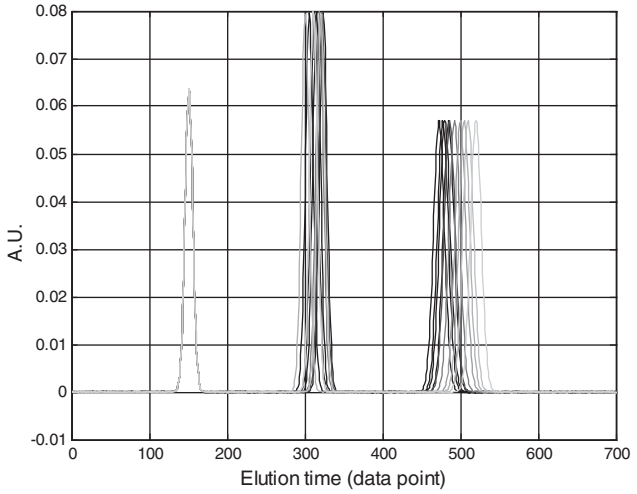


Figure 1. Illustration of 10 simulated chromatograms, each containing three Gaussian peaks. First peak (left): no shifts, second peak (middle): random shifts, and third peak (right): systematic peak shifts where a higher sample number is related to a later eluting time. Low intensity additive random noise has been added to all 10 chromatographic profiles, in both peak and baseline regions.

The method presented is valid for data sets with rather homogenous samples with similar chromatographic profiles (see Figures 1–3 for examples). However, more complex sample sets containing, for example, missing peaks across the samples or samples from two chromatographic column of different length will most probably need some pre-alignment before the presented method can be applied.

2. THEORY

2.1. Nomenclature and terminology

All measurement vectors will be referred to as sample chromatograms or simply *chromatograms*, independent of the analytical technique used. The direction along which the chemical constituents elute and where warping/alignment is required is referred to as *time*. Throughout this work, lowercase italics are used for scalars (i.e. x) and lowercase bold for row vectors (i.e. $\mathbf{x}$)—for example, a chromatographic profile. T in superscript is the transpose operation (i.e. $\mathbf{x}^{\mathrm{T}}$ is a column vector). Data matrices will be denoted with bold capital letters (i.e. $\mathbf{X}$). The ijth element of $\mathbf{X}$ is thus denoted $x(i,j)$, where the indices run as $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

2.2. Correlation optimised warping (COW)

The COW algorithm was introduced by Nielsen *et al.* [1] as a method to correct for shifts in discrete data signals. It is a piecewise or segmented data preprocessing technique that uses Dynamic Programming to align a sample chromatogram towards a reference chromatogram by stretching or compression of sample segments using linear interpolation.

There are at least two different ways of implementing COW [8]. This paper uses a slightly modified version of the COW algorithm developed by Tomasi *et al.* [3], which uses the summed correlation coefficient as optimisation criterion for determining the optimal path for alignment (largest value of the summed correlation coefficients). Apart from some computational aspects briefly discussed below, the main change with respect to the original algorithm by Tomasi *et al.* [3] concerns the boundaries between adjacent segments. Specifically, in the newer version used here, adjacent segments share the boundary they have in common, whereas in the old one, the boundaries of adjacent segments were two distinct and consecutive points [3]. Since the latter choice represented an inconsistency with the original algorithm by Nielsen *et al.* [1] and may entail an increased number of discontinuities in smooth signals whenever large corrections are allowed [8], it was removed in the implementation used herein.

Other modifications concerned strictly computational aspects and were aimed at reducing the computation time by reducing the number of operations and by increasing code vectorisation and recourse to built-in functions. In particular, three measures provided the largest reduction in time consumption: reduction in the number of operations for the (a) interpolation step, (b) for the computation of the correlation coefficient, (c) operation on all samples instead of one at a time. In particular, in the new fast version, the

DOI: 10.1002/cem

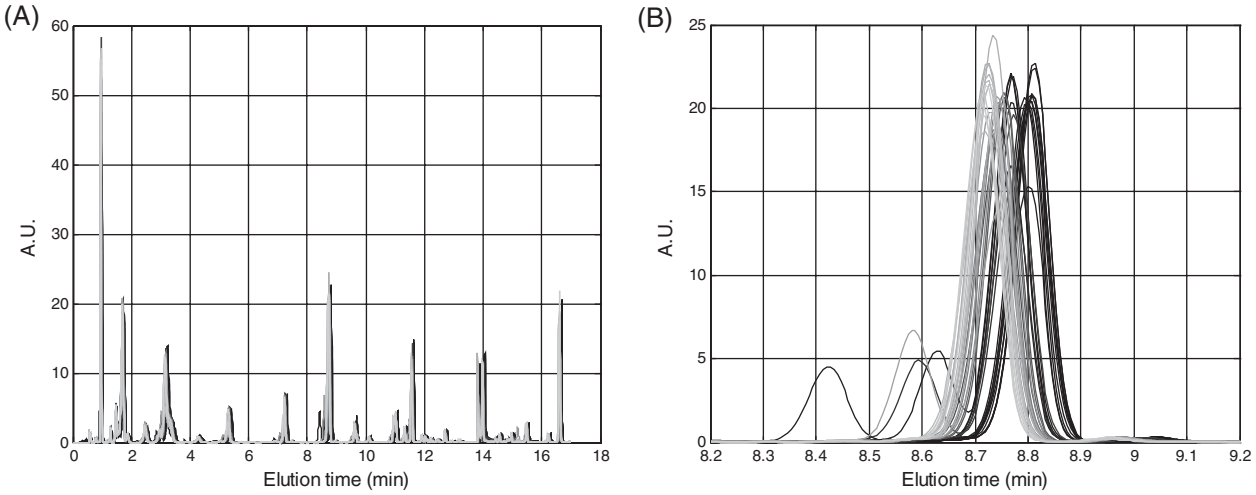


Figure 2. (A) Illustration of the shift problem in gas chromatograms (GC) with flame ionisation detection (FID) for grounded coffee extracts. (B) A zoom-in of the elution time 8.2–9.2 min.

correlation coefficient between two vectors $\mathbf{x}$ and $\mathbf{y}$ of length N is calculated as:

$$r(\mathbf{x},\mathbf{y}) \equiv \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}} = \frac{\left[(\mathbf{I}_N - \mathbf{1}\mathbf{1}^T N^{-1})\mathbf{x}\right]^T (\mathbf{I}_N - \mathbf{1}\mathbf{1}^T N^{-1})\mathbf{y}}{\|\tilde{\mathbf{x}}\|_2 \left(\|\mathbf{y}\|_2^2 - N\overline{y}\right)^{1/2}} = \frac{\tilde{\mathbf{x}}^T \mathbf{y}}{\|\tilde{\mathbf{x}}\|_2 \left(\|\mathbf{y}\|^2 - N\overline{y}\right)^{1/2}} \quad (1)$$

where $\tilde{\mathbf{x}}$ denotes the centred $\mathbf{x}$, $\overline{y}$ is the mean of $\mathbf{y}$ and the last equivalence is due to the fact that the centring matrix $\mathbf{I}_N - \mathbf{1}\mathbf{1}^{\mathrm{T}} N^{-1}$ is symmetric and idempotent. Note that using Equation (1) it is not necessary to centre $\mathbf{y}$ to compute r and that, if $\mathbf{x}$ is taken as a segment in the reference, its norm and centred form are identical for all samples and all possible boundaries for the corresponding sample segment. Likewise, the interpolated values are calculated using the expression: $x'(j) = x(l_j) + \alpha_j \dot{x}(l_j)$ where $x'(j)$ is the jth value of the interpolated sample segment, l_j is an index that depends only on j and on the segment length, $\dot{x}(j) = x(j+1) - x(j)$, and $\alpha_j \in [0,1]$ depends on the segment length and on the number of points in the interpolated samples. As can be seen indexes and coefficients do not depend on the sample and, like $\dot{\mathbf{X}}$, are calculated beforehand in the modified COW. The theory of COW will not be explained further here but the reader is referred to the literature for more details [1,3,4,8].

2.3. Reference chromatogram selection

The selection of reference sample (i.e. reference chromatogram) is often made from *a priori* knowledge on the data set. This could, for example, be the chromatogram in the middle of the run sequence [2,10] or the chromatogram containing the highest number of common chemical constituents (i.e. peaks) [3–4]. However, to make sure that the most

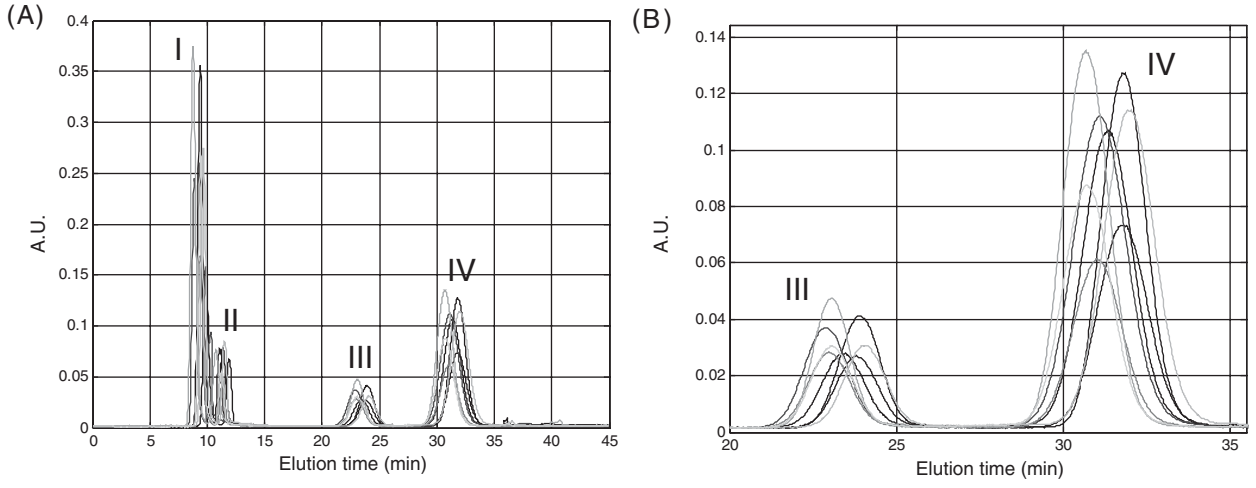


Figure 3. (A) Illustration of the shift problem in high performance/pressure liquid chromatography (HPLC) with fluorescence detection of four different isomers (I: α-Tocopherol, II: α-Tocotrienol, III: β-Tocopherol, IV: β-Tocotrienol) of Vitamin E in wheat flour samples. (B) A zoom-in of the elution time 20.0–35.5 min.

DOI: 10.1002/cem

appropriate reference sample is selected in a given data set, a more objective method is needed. One solution could be to choose the chromatogram that is most similar to the loading of the first principal component in a PCA model on the unaligned data or simply to the mean of all chromatograms. Such a generic approach can be problematic because the mean chromatogram as well as the first loading from a PCA of the raw data will have too many or heavily distorted/broadened peaks due to the original problem at hand: the shifts present in the data set.

In this paper, a method is presented which is based on the product of the correlation coefficients between all individual chromatograms. For a given chromatogram $\mathbf{x}_t$, this *similarity index* ($0 <$ similarity index ≤ 1) can be calculated as:

$$\text{Similarity index} = \prod_{i=1}^{I} |r(\mathbf{x}_t, \mathbf{x}_i)| \tag{2}$$

where $r(\mathbf{x}_t,\mathbf{x}_i)$ is the conventional correlation coefficient between two chromatograms in the data set calculated as shown in Equation (1).

Taking the absolute value in Equation (2) will safeguard the similarity index selection from the situation where strongly deviating samples in the data set will have a low correlation coefficient with arbitrary sign. However, like in all data processing operations, such samples should preferably be caught and removed before further computations. The similarity index for each sample in the set will be less then or equal to one (in case of perfectly aligned and identical chromatograms). The chromatogram that is most similar to all others will have the largest similarity index and is selected to be the most suitable reference chromatogram to use within the given data set.

2.4. The simplicity value

The overall goal when aligning chromatograms is to make the profiles as similar in appearance as possible while preserving the peak shape and area. Stated differently, with the right preprocessing, the numerical rank of the data set, disregarding random noise, will be lowered towards the chemical rank [3].

The *simplicity* value is used to measure how well aligned a set of chromatograms is. The principle of the simplicity value is related to the properties of the singular value decomposition (SVD), where the size of the squared singular values is directly related to the sum of squares of the data matrix. Any data matrix, $\mathbf{X}$ (uncentered) can be decomposed as:

$$\mathbf{X} = \mathbf{USV}^T \tag{3}$$

where $\mathbf{S}$ is a diagonal matrix containing the singular values equal to the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal matrices, where the columns in $\mathbf{U}$ are the eigenvectors of $\mathbf{XX}^T$ and the columns of $\mathbf{V}$ the eigenvectors of $\mathbf{X}^T\mathbf{X}$. The sum of the squared singular values equals the total sum of squares of all the original data entries in $\mathbf{X}$. Thus, the ratio of each individual squared singular value divided by the sum of all squared singular values can be regarded as a measure of how much of the variation in $\mathbf{X}$ is partitioned into each component.

In unaligned data the chromatographic profile will differ among samples and thus less of the total variation will be explained by the first few singular values. This is due to the deviation from low-rank bi-linearity, which causes more significant singular values in the decomposition of the data matrix. If data are aligned and all chromatographic profiles only differ in the magnitude from a common profile, then the above ratio for the first singular values will be equal to one. Generally, better-aligned chromatograms will result in fewer and larger significant singular values, which would represent the (true) chemical information we are looking for.

The sum of the first R squared singular values is a measure of how much of the variation is explained by the corresponding R components:

$$\text{Explained variance} = \sum_{r=1}^{R} \left(\text{SVD}\left(\mathbf{X} / \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} x(i,j)^2} \right) \right)^2 \tag{4}$$

where SVD($\mathbf{M}$) denotes the singular value for a given component r and where the data are scaled to a total sum of squares of one for convenience. The above expression is by definition equal to one if all singular values are retained, and as such this sum cannot be used to evaluate preprocessing and the effect of alignment. However, for finding the optimal combination of segment and slack size the following expression is put forward as the *simplicity* value ($0 \leq$ simplicity ≤ 1), where the principle of simplicity is adapted from Henrion & Andersson [9], Christensen *et al.* [10] and Johnson *et al.* [11]:

$$\text{Simplicity} = \sum_{r=1}^{R} \left(\text{SVD}\left(\mathbf{X} / \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} x(i,j)^2} \right) \right)^4 \tag{5}$$

This sum of the singular values taken to the fourth power will be higher the more of the variation that is explained by the first components. As a simple example consider the two alternative series of singular values $\text{SVD}(\text{M}_1) = [1\,0\,0\,0]^\text{T}$ and $\text{SVD}(\text{M}_2) = [\tfrac{1}{2}\ \tfrac{1}{2}\ \tfrac{1}{2}\ \tfrac{1}{2}]^\text{T}$. Although the squared sum of the series is the same, the fourth-power sum is 1 and 1/4, respectively, indicating that in the first case, the data are more similar in shape as seen from the lower rank; they can be explained with one component only. In general the simplicity value will be noticeably smaller if the chromatograms are not well aligned. Having achieved perfect alignment the simplicity value will be closer to though not necessarily equal to one.

In COW alignment, it is often possible to achieve high simplicity values with several combinations of segment and slack parameters. This is illustrated in Figure 4(A) for the simulated chromatograms. It is obvious that combining a small segment length with a large slack size (i.e. high flexibility) will result in interpolation steps over many data point and thus the possibility to align peaks efficiently, but this also carries the danger to undesirably change both shape and area of peaks. It should be emphasised that the presented method focuses on preserving the total area of all peaks in the chromatographic profiles, stating that any change introduced by the alignment procedure is undesirable. To avoid this potential pitfall, we include a second criterion in the optimisation of simplicity that takes into account this 'change in area' effect and can guide the selection of the optimal

DOI: 10.1002/cem

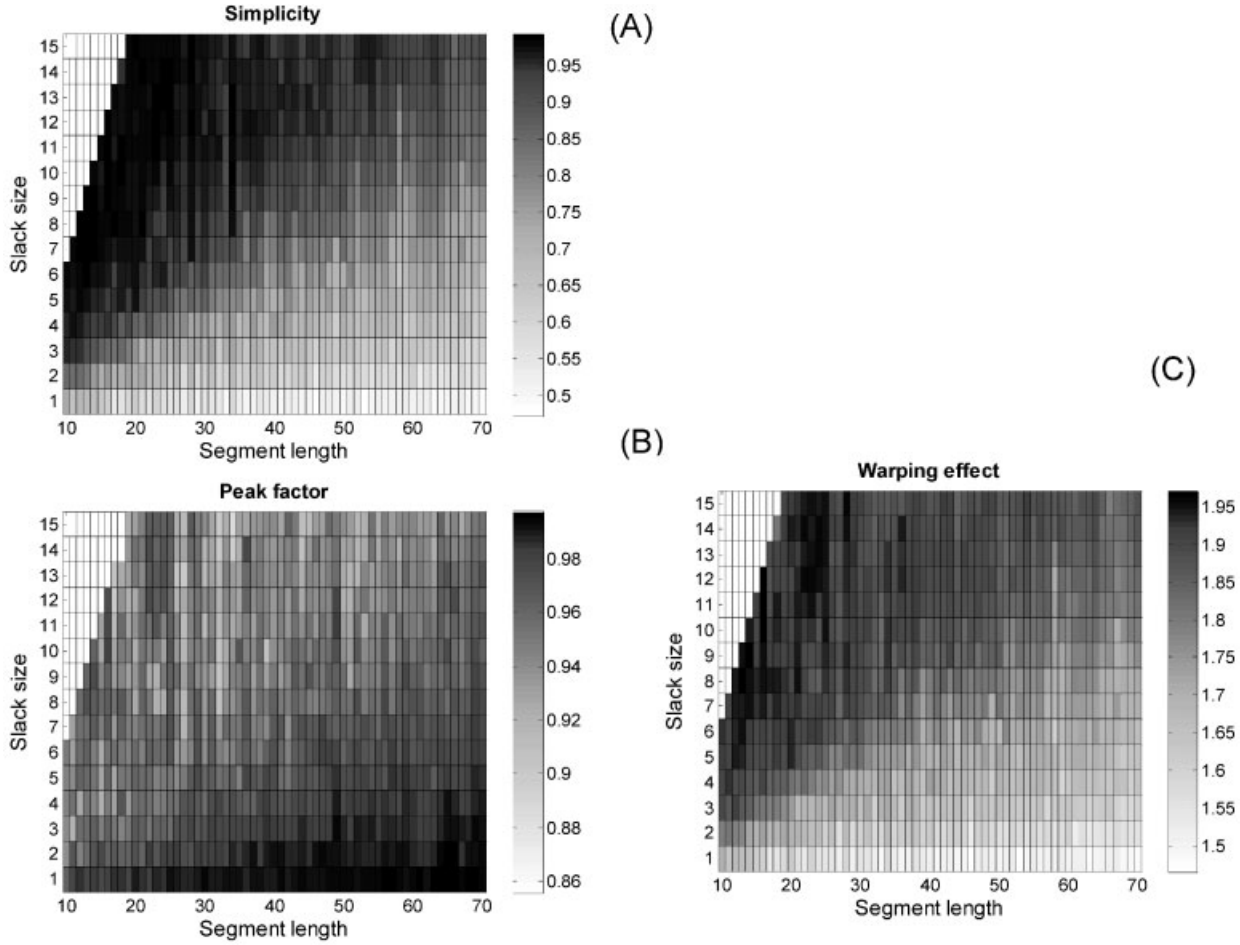


Figure 4. Simplicity (A), peak factor (B) and warping effect (C) values for all combinations of segment length and slack size using the simulated data. For plots (A) and (B) a value close to one indicates that data are well aligned and that the area has changed insignificantly, respectively. For plot (C) a value close to two means that peaks are both aligned and that the change in the area is minimal. The white triangle in the upper left corner contains unfeasible combinations of segment length and slack size in the COW algorithm.

combination of segment and slack via an additional penalty term.

2.5. The peak factor

When aligning chromatograms, the peak area and shape should ideally be the same before and after the procedure; where all required corrections should ideally take place in the baseline area of the profile. A prerequisite for this ideal situation is that the data should be rather homogenous to make sure that no very strong compression (or stretching) is required which will change the shape of peaks and thus affect the peak area.

One prerequisite for a successful preprocessing is that the reference chromatogram has been carefully selected, but this alone cannot guarantee that peak shapes and areas do not change. We quantify the change by a measure called *peak factor* ($0 \leq$ peak factor ≤ 1). It indicates how much the total sample set has changed when preprocessed by a particular combination of segment length and slack size:

$$\text{Peak factor} = \frac{\sum_{i=1}^{I} \left(1 - \text{minimum}(c(i), 1)^2\right)}{I} \tag{6}$$

where,

$$c(i) = \left| \frac{\|\mathbf{x}_w(i)\| - \|\mathbf{x}(i)\|}{\|\mathbf{x}(i)\|} \right| \tag{7}$$

and $\|\mathbf{x}(i)\| = \sqrt{\sum_{j=1}^{J} x(i,j)^2}$ is the Euclidian length or norm for $\mathbf{x}$(i); $\mathbf{x}$(i) is the chromatogram before warping while $\mathbf{x}_w$(i) is the same sample after alignment. In this criterion (7), if the norm stays the same, the absolute term (or relative change) is 0, and the overall contribution for that sample is 1 in Equation (6). If a sample is almost the same, the absolute term will be between 0 and 1, and the overall contribution will be smaller than 1. When the warped sample is very distorted, the absolute term will grow, and its overall contribution will be 0. By using the Euclidian length, larger peaks will be relatively more influential for the peak factor. This choice works well for the type of signals presented in this study, but for different data analytical problems (e.g. investigations on trace amounts hidden in a complex, dominant matrix) alternative norms might be more appropriate.

Values of the peak factor measure are shown in Figure 4(B) together with the simplicity values for the simulated data. Notice that some combinations of segment length and slack

DOI: 10.1002/cem

size provide high simplicity values but 'low' peak factor values, and thus should not be considered as suitable alignment parameters.

2.6. The warping effect

The new quantitative measure combining the simplicity and the peak factor value is called the *warping effect* ($0 \leq$ warping effect ≤ 2):

$$\text{Warping effect} = \text{simplicity} + \text{peak factor} \tag{8}$$

The relation between the three measures—simplicity, peak factor and warping effect—is illustrated in Figure 4 for the simulated chromatograms.

Figure 4 shows that some combinations of segment length and slack size result in a potential distortion of the profiles and thus are inappropriate choices for the alignment parameters. The warping effect overall gives a smoother, less pronounced optimisation landscape with more gradual changes as compared to the simplicity plot, opening the opportunity for an automated optimisation routine which will have a better chance of finding (near) optimal parameters of segment length and slack size in reduced computation time.

Equation (8) is simply the sum of the two different criteria presented earlier. No separate weight factors are introduced and both terms thus have the same influence on the warping effect value. This works satisfactory for homogenous data (i.e. the three examples shown in this paper) where the alignment term (simplicity) will typically have a higher influence, as the difference between good and bad alignment values is larger (in some situations it will be only slightly larger) than good and bad area preservation values (Figure 4(A,B). Alternative strategies for relative importance of the two terms are possible, but this subject will not be pursued in this paper.

2.7. Optimisation

An exhaustive search for the optimal combination of segment length and slack size will be rather time consuming for realistic problems, and as illustrated in Figure 4 is not rational as (near) optimal solution(s) (i.e. combinations) are present over a considerable, well defined area/corner of all segment length and slack size combinations. For the situation depicted in Figure 4 an exhaustive search involves the calculation of $61 \times 15 = 1065$ combinations. Many of these combination-areas are suboptimal, and thus, it would be more rational to limit the computational cost by finding a good starting point for further optimisation. One way this can be achieved is by using a predefined sparse search grid, equally spaced over the total search region, as illustrated in Figure 5 for the simulated chromatograms.

The optimisation of the warping effect values is done in the form of a discrete-coordinates simplex-like optimisation routine carried out in three steps [12]. First step in our optimisation is to establish global search space boundaries from the combination of all segment length and slack sizes of interest. In the second step, a sparse global search grid is determined where we by default select a 5×5 grid in both the segment and slack direction, as indicated in Figure 5, and the warping effect for these 25 points is determined. The six best

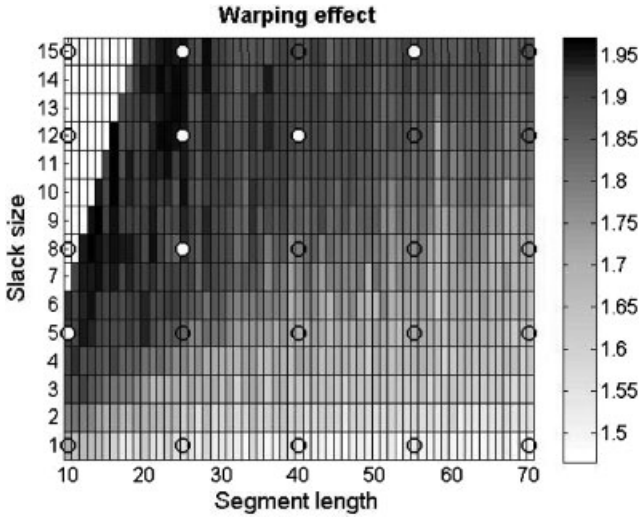


Figure 5. 5×5 sparse search grid (circles) for the global search space of segment length 10–70 and slack size 1–15 using the simulated data as an example. The six largest 'winning' parameter sets of the warping effect in this search grid are indicated with solid dots.

(default choice) combinations, providing the highest warping effect scores, are then selected and used as starting points in a discrete-coordinates simplex optimisation part. This routine is depicted in Figure 6. It works by establishing a triangle with the base (the intersection of the two legs of length one) on the grid point (triangle I). The three points (combinations of segment and slack) are then calculated for as far as they were not known *a priori*, and the triangle is flipped over the side formed by the two points possessing the largest warping effect (I to II). If the value in the corner of the new triangle after evaluation is lower than the one found in the previous step the flip is not continued (II to III). The

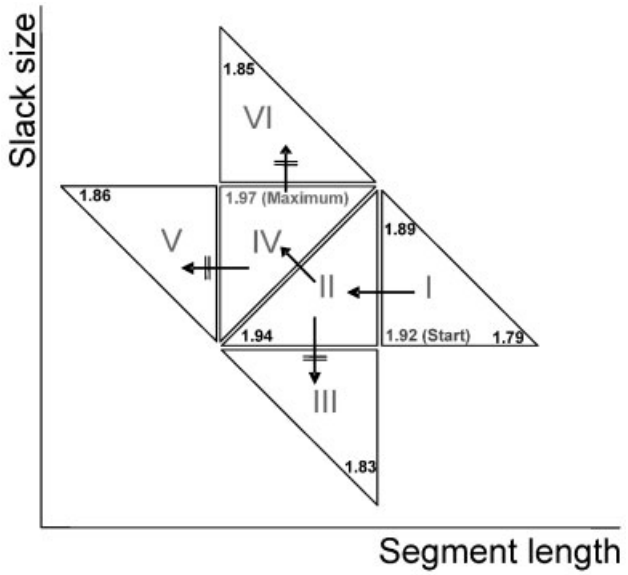


Figure 6. Illustration of the discrete-coordinates simplex-style optimisation routine using warping effect values. See text for explanation of optimisation steps. The roman letters indicate the consecutive steps of the optimisation routine.

DOI: 10.1002/cem

Table 1. Optimisation of warping effect values: six starting points for the discrete-coordinates simplex optimisation routine to find the optimal alignment parameters for the simulated chromatograms.

Optimisation	Starting point (segment/slack)	Warping effect	Optimised to (segment/slack)	Steps	Warping effect	Simplicity value	Peak factor
1	25/15	1.9486	24/13	9	1.9614	0.9890	0.9724
2	25/12	1.9433	24/13	8	1.9614	0.9890	0.9724
3	25/8	1.9248	25/9	4	1.9373	0.9746	0.9627
4	10/5	1.9179	9/5	5	1.9410	0.9790	0.9619
5	40/12	1.8879	38/11	12	1.9227	0.9652	0.9575
6	55/15	1.8816	56/17	8	1.9129	0.9638	0.9491

routine instead flips the triangle over the second highest side and makes a new evaluation (II to IV). The optimisation stops when no further flips are possible according to the rules explained above and visualised in Figure 6.

The main characteristics of the optimisation routine for the simulated data are highlighted in Table 1. The end point in the path of each of the six optimisation steps can be seen in Figure 7.

As seen in Table 1 the best combination of segment length and slack size (24/13) holds the largest simplicity and peak factor value of the six end points. However, for real data the situation is often more complex as the warping effect value is a compromise between alignment and peak area preservation. For some experiments the alignment might be most important and in others the preservation of the peak area more essential.

The optimisation space (and sparse search grid) in Figure 7 includes combinations of segment length and slack size close to the dark area located in the upper left corner. This means that only a few steps are required for the simplex-style optimisation routine to find these combinations assuming that no local maxima are found along the path. The distance in optimisation steps to (near) optimal solutions and the smoothness of the global search space are the two main issues that affect the optimisation routine. The first issue is taken care of using a predefined sparse search grid as shown in Figure 5, whereas the other part is accomplished by using the warping effect instead of simplicity values in the optimisation routine.

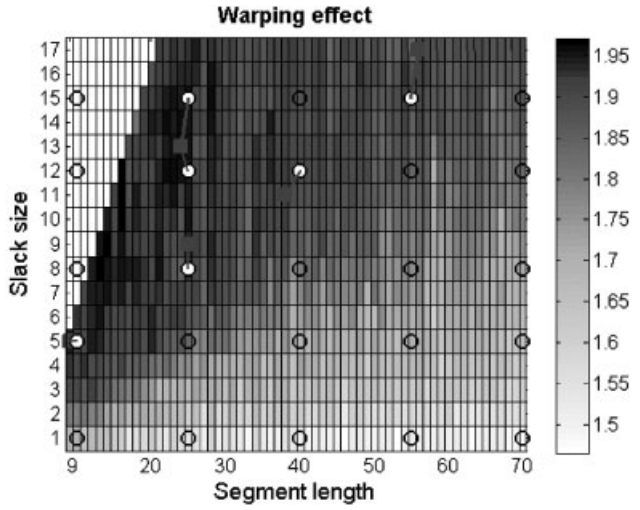


Figure 7. Warping effect: start points (filled circles) and end points (filled squares) in the optimisation path for each of the six starting values. Notice that in contrast to Figures 4 and 5, the slack dimension had been extended to include size 17 as one of the optimisation end points include this slack size. The time consumption for the optimisation routine is 1.5 min, whereas the exhaustive search (to find the global maximum value) takes 15 min. The ratio between these times depends on the size of the search space and the steps required by the optimisation routine, as will be shown in Figures 10 and 12 for the GC and HPLC data, respectively.

For the warping effect values in Figures 4, 5 and 7, darker regions can be observed; one of them holding the maximum of the entire search space. However, local combinations can give warping effect values just as high, but they can be difficult to detect in this kind of graphical presentation. Thus, the use of multiple starting points on the sparse search grid can lead the optimisation paths in other directions, as will be shown for the GC and HPLC data. The overall routine still ends up in combinations of segment length and slack size with near optimal alignment characteristics.

To illustrate the effect of including the change in area (the peak factor) in the optimisation routine, three examples of aligned peaks are shown in Figure 8. The examples are the best combination found from the optimisation routine, the worst starting point and the second best solution, which can all be found in Table 1. Notice that all peaks in this area of the simulated data are the same Gaussian peak and any (undesirable) deviations from this shape is due to the alignment preprocessing.

Figure 8 shows that even though the samples in the first peak region (from 120 to 180 data points, see Figure 1) are perfectly aligned in the raw data, small changes can be introduced as a result of the alignment process because it aligns the total chromatogram which is heavily dominated by the last peak. All three combinations give simplicity values that would indicate an efficient alignment, but the change in area results in the combination of segment length 24 and slack size 13 as the outcome of the optimisation routine. For this combination the change in shape of the shown peak is also found to be insignificant, whereas the other combinations result in some deviation from the original peak shape.

2.8. Defining the optimisation space

As shown in Figure 4, the search space for the segment lengths includes several feasible choices as long as the

DOI: 10.1002/cem

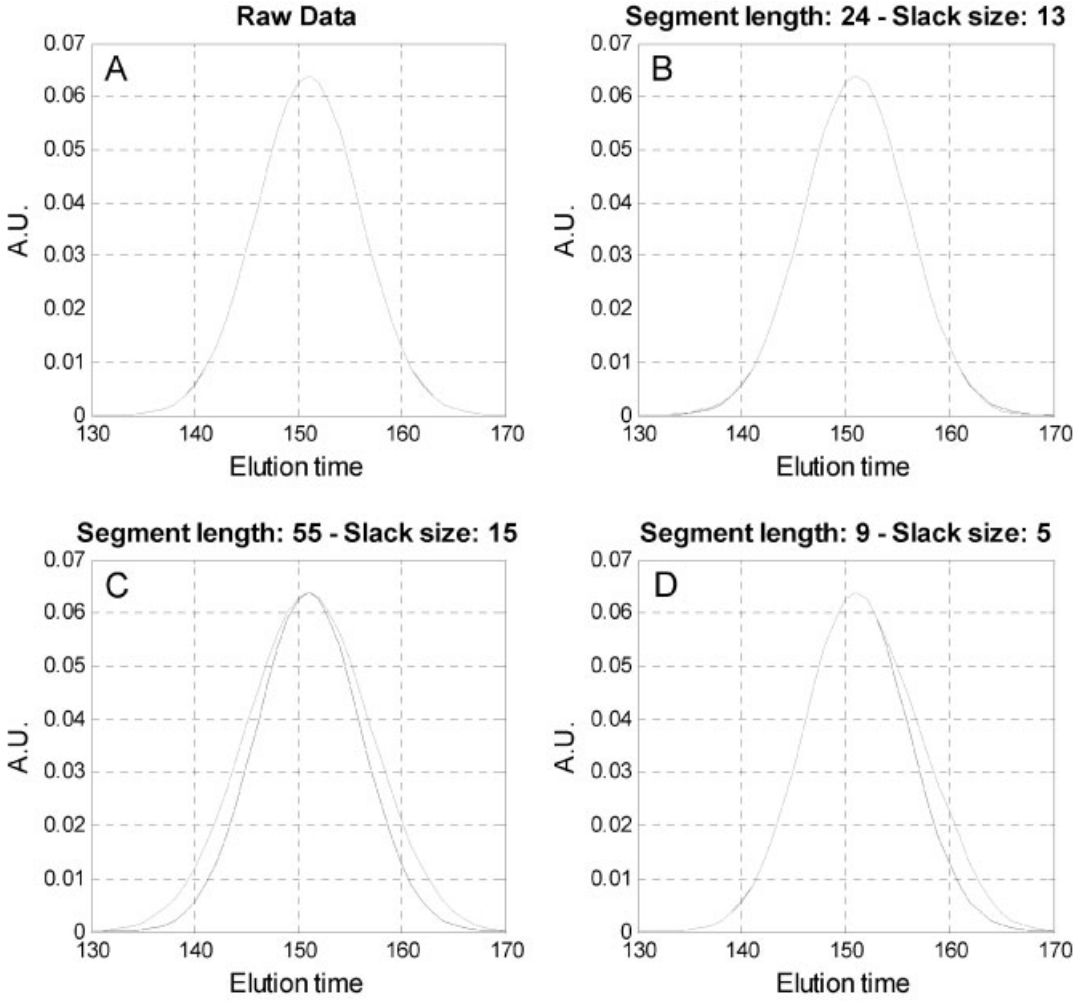


Figure 8. Illustrations of alignment and peak area characteristics for key combinations of segment length and slack size for two samples in the simulated chromatograms on single peak eluting between 130 and 170 data points. (A) Unaligned, (B) optimised to segment length: 24, slack size: 13, (C) worst starting point: segment length: 55, slack size: 15 and (D) second best optimised solution: segment length: 9, slack size: 5. For the combinations shown; an average change in the total sample set (calculated as (1-peak factor) $\times$ 100) of 2.8%, 7.7% and 3.8%, respectively, is found.

flexibility (slack size) is large enough. In general, longer segment lengths require more flexibility to give good alignment. However, this depends on the chromatographic data at hand so the following guidelines should be used with some care. We still assume that the chromatographic profiles are rather homogeneous and as such the following guidelines might not hold when dealing with more complex profiles, for example, very shifted data, as explained in the introduction.

2.8.1. *Segment length*

A rule of thumb for selecting the segment length optimisation space is:

$$PW_A \pm \frac{PW_A}{2}$$

where PW_A is the approximate peak width average at the base over all peaks in the reference chromatogram. By this rule, the segment lengths will contain both peak fragments and entire peaks. In Section 3, this rule of thumb will be used to set the upper limit for the segment length, whereas the lower limit is set to 10 points for all calculations. This is done here mainly for illustrative purposes, to show the effect of combining a low segment length and a high flexibility (large slack size) on the preservation of the peak areas.

2.8.2. *Slack size*

The right slack size search space is more difficult to define as features such as different local peak shifts, data points before and after the first and last peak, and increased flexibility of the COW algorithm in the middle of the chromatogram will have an effect on the outcome of the alignment procedure. However, a rule of thumb is that if the number of data points before and after the first and last peak, respectively, is approximately the same as the peak widths (ensuring enough flexibility), then a slack size search space ranging from 1 to 15 is appropriate, also considering that higher values will increase the computation cost considerably.

DOI: 10.1002/cem

Effective slack size depends on the chromatographic technique used, but it has been found that mostly setting the space as above provides reasonable results. The outcome of an optimisation may indicate that the slack size search space needs to be adjusted if only the extreme slacks are chosen. In the following the lower limit will always be set to one, whereas the upper limit is set higher than the suggested if peaks are more shifted (e.g. HPLC data) and lower when only small shifts (e.g. GC data) are observed.

3. RESULTS AND DISCUSSION

3.1. Coffee data

The GC-FID data, shown in Figure 2, was presented before by Tomasi *et al.* [3] and is used in this study as the first example. It was obtained by gas chromatography (GC) analysis of extracts from ground coffee according to the experimental conditions described in [3]. The number of samples in the data set used for the optimisation routine has been reduced by taking every second object in the original coffee data set giving 42 samples $\times$ 2550 elution times. The optimal reference would be a chromatogram that includes all local (sample-wise) variations (peaks). As this is rarely achievable, the best alternative is to find the most suitable reference from within the given data. Our approach has been given in the Theory Section of this paper, which selects the sample which is most similar to all others based on the product of correlations coefficients. This similarity index calculated for the samples in the coffee data set can be seen in Figure 9.

As can be seen from Figure 9, many samples give small similarity indexes. This is an effect of the rather severe down-weighting by multiplication of the product of correlation coefficients which might give the impression that samples are very different. It can also be seen that samples in the middle of the run sequence are most suitable as reference samples, which was expected as the shift effect and the time of measurement are confounded for this data set [3]. Thus, having no proper reference selection tool one might select any sample in the middle of the run sequence based on a visual inspection of raw chromatograms. However, some samples in the middle give low similarity indexes (e.g. sample #25) and thus an objective reference selection tool is useful and can replace a time consuming and possibly error prone visual method.

3.2. Alignment parameters—initial considerations for real data

As explained previously, the main goal of introducing the simplicity parameter is to get an objective measure of how well the chromatograms have been aligned towards the chosen reference chromatogram. The selection of the reference sample will obviously affect the proper combinations of segment and slack. Selecting a reference sample in the beginning of a chromatographic run (e.g. sample # 1 in Figure 2) will require the last sample in the run to be shifted significantly more towards the reference. Thus, a higher flexibility (larger slack size) is needed and together with that a possibility for significant changes of peak area through the interpolation steps.

The peak shape plays a major role in the alignment procedure. Overloading a chromatographic column can cause peaks to tail or front in such a way that the peak shape changes significantly, whereas the area is kept proportional to the concentration of the eluting compound. Aligning peaks of different shapes can cause a dramatic change in the area of peaks of the sample chromatogram, as the interpolation of data points will be guided by the shape of peaks in the reference chromatogram. The COW algorithm is best suitable for chromatographic data holding peaks that do not change shape across chromatograms.

For the coffee data, the global search space is set to be a combination of segment length from 10 to 80 (average peak width is around 45 data points) and a slack size from 1 to 10 (small shifts); the results are presented in Figure 10.

In Figure 10, the six starting points determined from a 5 $\times$ 5 sparse search grid for the optimisation routine on the coffee

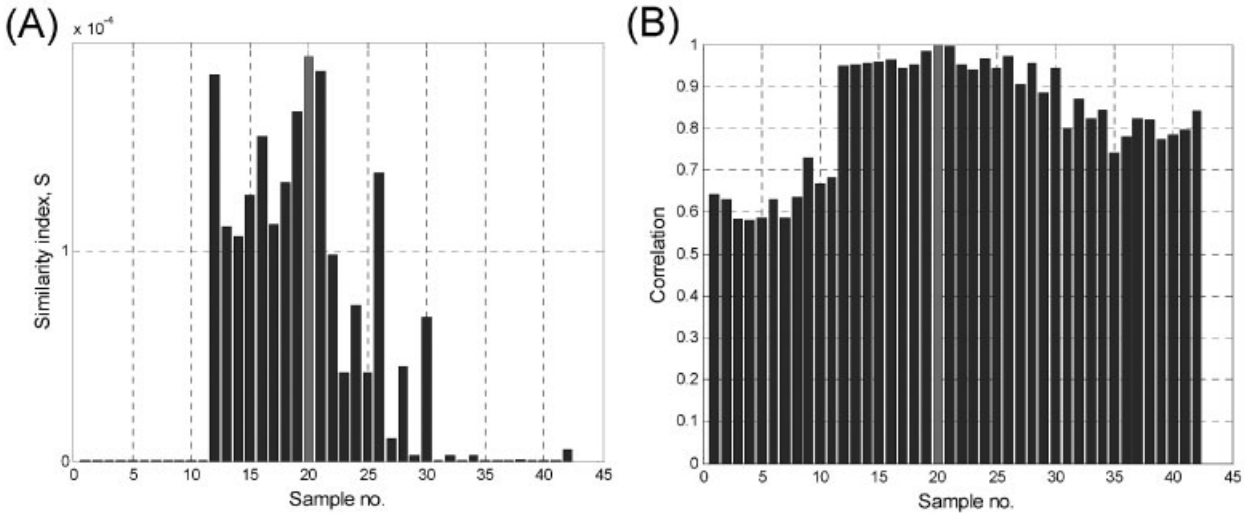


Figure 9. (A) Similarity index from Equation (2) calculated for all samples in the coffee data set. The sample having the highest similarity index (here sample #20) is selected as the reference sample during the alignment procedure. (B) The correlation coefficient from Equation (1) between reference sample #20 and all other samples in the set.

DOI: 10.1002/cem

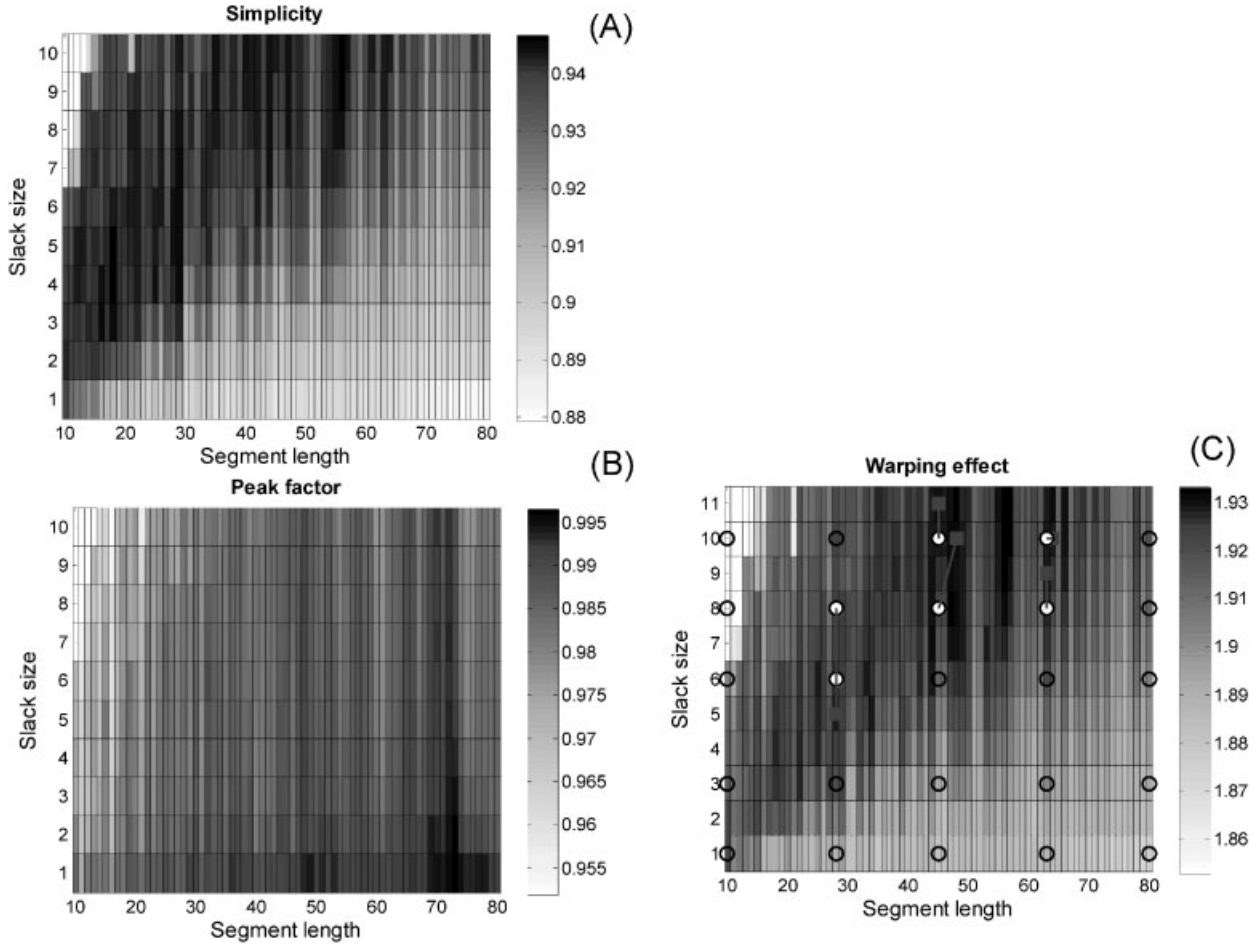


Figure 10. (A) Simplicity, (B) peak factor and (C) warping effect values for all combinations of segment length and slack size using the coffee data. Warping effect: start points (filled circles) and end points (filled squares) in the optimisation path for each of the six starting points. Further information can be found in Table 2. The 5×5 sparse search grid is visualised in (C). The time consumption for the optimisation routine is 23 min, whereas the exhaustive search takes 12.2 h.

data are marked (results summarised in Table 2). Using only the simplicity value in the optimisation routine, these points would primarily be found in the lower segment length and higher slack size region, as explained earlier. However, including the peak factor this region becomes less dominant, as severe changes in shape and area also occur there due to combination of shorter segments and increased flexibility. Figure 10 also shows that the difference between best and worst alignment (simplicity value) and change in area are relatively small and thus all combinations will to some degree improve the alignment. This is mainly because this particular chromatographic data has a systematic peak shift over the entire elution time and thus small alignment improvements and detrimental area changes can be difficult to detect in the overall picture. In each of the six grid points the discrete-coordinates simplex optimisation routine is carried out and a final combination of segment length and slack size is found (Table 2). The end point in the path of each of the six optimisation steps can be seen in Figure 10(C).

Table 2. Optimisation from warping effect values: six starting points for the discrete-coordinates simplex optimisation routine to find the optimal alignment parameters for the coffee data

Optimisation	Starting point (segment/slack)	Warping effect	Optimised to (segment/slack)	Steps	Warping effect	Simplicity value	Peak factor
1	45/10	1.9293	45/11	5	1.9305	0.9440	0.9865
2	63/10	1.9284	64/10	5	1.9292	0.9427	0.9865
3	63/8	1.9275	63/9	4	1.9290	0.9435	0.9854
4	28/6	1.9274	28/5	5	1.9282	0.9442	0.9841
5	28/8	1.9248	28/5	11	1.9282	0.9442	0.9841
6	45/8	1.9234	48/10	9	1.9314	0.9409	0.9905

DOI: 10.1002/cem

From Figure 10(C), it can be seen that the selected combinations from the sparse search grid are not far away from the dark regions holding the optimal combinations. Hence, in this case there is a good chance for the 5×5 grid part of the optimisation routine to end up in these regions. A more comprehensive optimisation could obviously be accomplished by defining a finer initial grid, but at the cost of a higher computational cost. This could be of relevance if a larger global search space was selected.

From Table 2 it is also noticeable that the optimisation path from the worst starting point ends up in the best-optimised solution. This is contradictory to what was observed for the simulated data where the best starting points also gave the best optimised solution. These two examples show the importance of the well-defined search grid illustrated in Figures 7 and 10.

In accordance with Figure 8, three combinations of segment length and slack size from Table 2 are plotted in Figure 11 for the GC data. Notice that only a small section of the total elution time profile has been plotted, but that all calculations in Table 2 and Figure 11 are based on the entire chromatogram

From Figure 11, it can be seen that all three combinations of segment length and slack size give well-aligned data and that peak shape changes are rather negligible. This could be explained by the systematic shift in the homogeneous GC data that gives the optimisation routine a good chance of finding one of the many near optimal combinations in the given search space.

Considering that these data consist of many peaks, the average change in area per peak is rather small, but must be taken into account if the subsequent objective is the use of the peak area after alignment in, for example, a factor model evaluation.

3.3. HPLC data

For the HPLC data the global search space is set to be a combination of segment length from 10 to 300 (average peak width of approximately 200 data points) and a slack size from

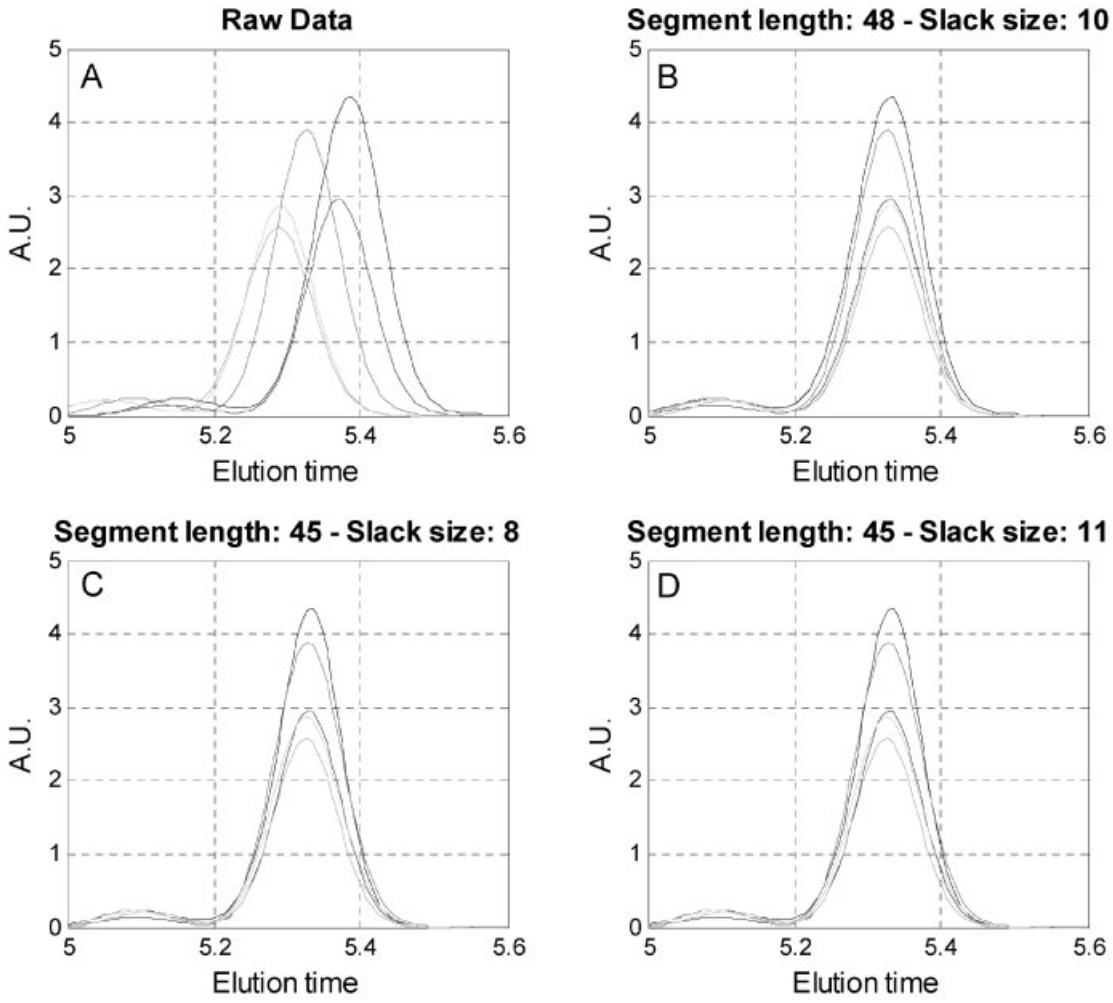


Figure 11. Illustrations of alignment and peak area characteristics for key combinations of segment length and slack size on peaks eluting between 5.0–5.6 min for five samples in the coffee data. (A) Unaligned, (B) optimised to segment length: 48, slack size: 10, (C) worst starting point: segment length: 45, slack size: 8 and (D) second best optimised solution: segment length: 45, slack size: 11. For the combinations shown; an average change in the total sample set (calculated as (1-peak factor) $\times$ 100) of 0.9%, 1.3% and 1.4%, respectively, is found.

DOI: 10.1002/cem

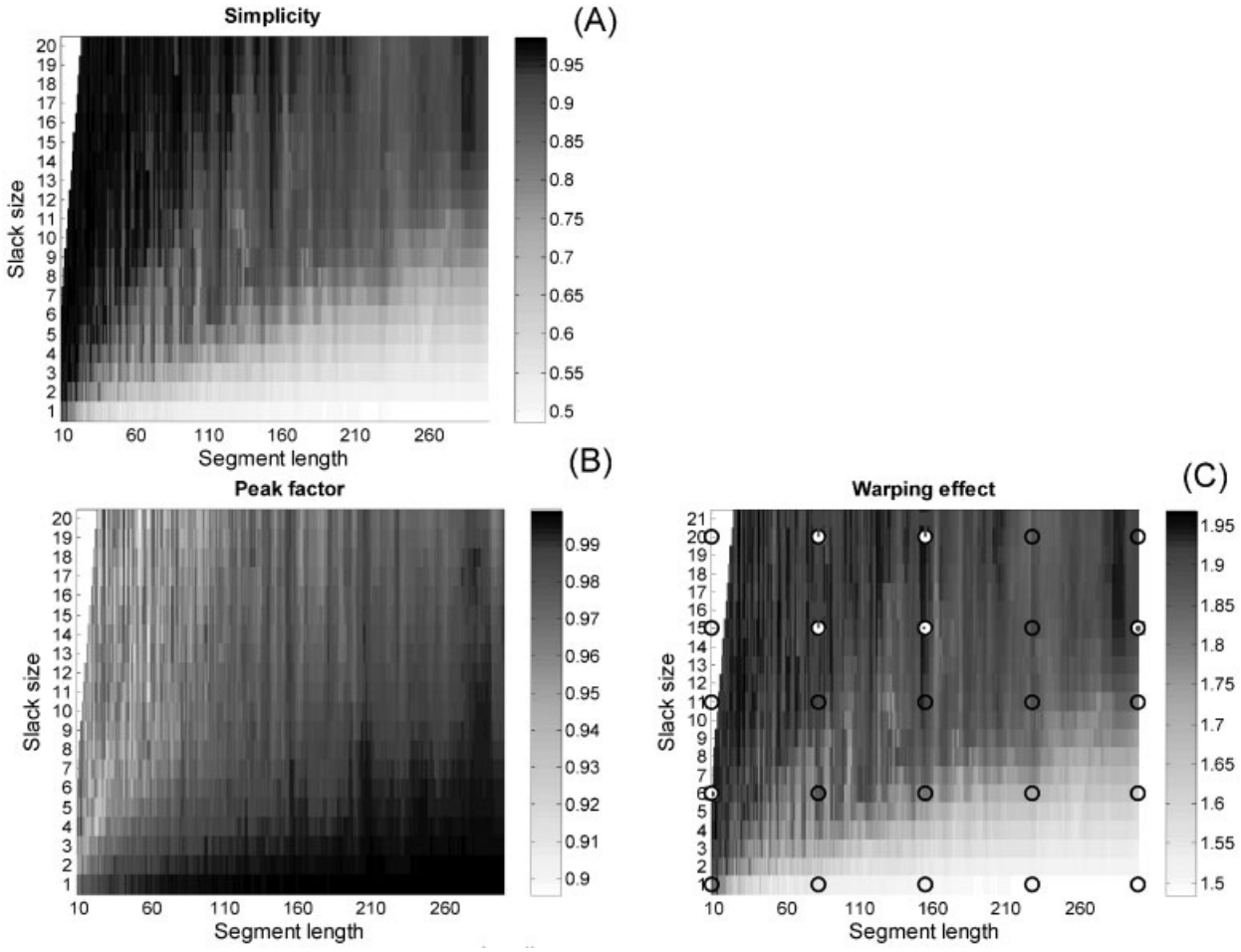


Figure 12. (A) Simplicity, (B) peak factor and (C) warping effect values for all combinations of segment length and slack size using the HPLC data. Warping effect: start points (filled circles) and end points (filled squares) in the optimisation path for each of the six starting points. Further information can be found in Table 3. The 5×5 sparse search grid is visualised in (C). The time consumption for the optimisation routine is 6.6 min, whereas the exhaustive search takes 5.8 h.

1 to 20 (severely shifted peaks). The quantitative measures for the alignment procedure are illustrated in Figure 12 together with the sparse search grid used in the optimisation routine. The details from the optimisation routine are summarised in Table 3 and the pathways from each starting points in this routine are illustrated in Figure 12(C).

In Figure 13, the key combinations of segment length and slack size that has been presented for the simulated and GC data are shown.

Two things are noticeable in Table 3 and Figure 13. Firstly, that the combination of segment length 10 and slack size 6 results in the second best overall optimised solution despite the severe changes in shape observed and secondly, that fewer steps are taken in all six optimisation paths. The first part confirms the relevance of introducing the peak factor in the overall warping effect expression. The second part shows that many local minima are present in the search space and that this optimisation might require a finer initial grid, to find

Table 3. Optimisation from warping effect values: six starting points for the discrete-coordinates simplex optimisation routine to find the optimal alignment parameters for the HPLC data

Optimisation	Starting point (segment/slack)	Warping effect	Optimised to (segment/slack)	Steps	Warping effect	Simplicity value	Peak factor
1	155/20	1.9564	155/21	5	1.9580	0.9696	0.9885
2	10/6	1.9458	10/6	5	1.9458	0.9777	0.9682
3	155/15	1.9368	154/15	5	1.9415	0.9549	0.9866
4	83/20	1.9229	83/21	4	1.9257	0.9664	0.9593
5	83/15	1.9111	83/16	4	1.9198	0.9540	0.9659
6	300/15	1.9085	300/15	5	1.9085	0.9189	0.9897

DOI: 10.1002/cem

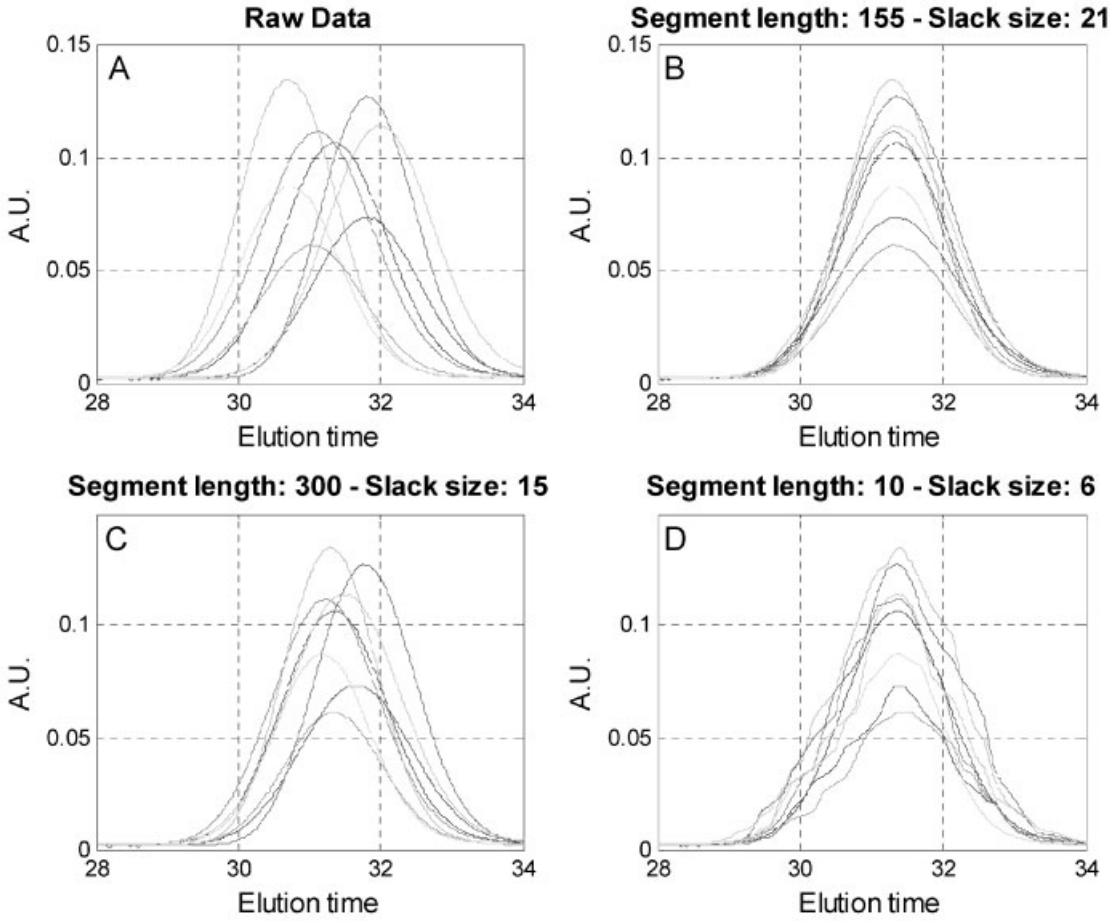


Figure 13. Illustrations of alignment and peak area characteristics for key combinations of segment length and slack size on single peak eluting between 28.0 and 34 min. (A) Unaligned, (B) optimised to segment length: 155, slack size: 21, (C) worst starting point: segment length: 300, slack size: 15 and (D) second best optimised solution: segment length: 10, slack size: 6. For the combinations shown; an average change in the total sample set (calculated as (1-peak factor) $\times$ 100) of 1.2%, 1.0% and 3.2%, respectively, is found.

an optimal combination. However, using the default 5×5 search grid an appropriate combination was still found demonstrating the effectiveness in the optimisation routine (Figure 6) and the well-defined search grid (Figure 5).

4. CONCLUSIONS

As shown with simulations and examples from different fields in chromatography (GC and HPLC), the effectiveness of the alignment of chromatographic profiles depends on factors such as reference sample, segment length and slack size. The selection of the reference sample will obviously affect the optimal combination of segment and slack. Selecting a reference sample in the beginning of a chromatographic series (e.g. first sample in Figure 2) will require the last sample in the run to be shifted significantly more towards the reference. Thus, a higher corrective flexibility is required (larger slack size) which comes with a possibility for significant changes of peak areas due to the interpolation steps. By selecting automatically the most suitable reference (with highest similarity index) the global search space can be significantly reduced and provide a significantly faster optimisation routine.

The determination of segment length and slack size search space can be made based on the rules of thumb given. These guidelines can be used if no knowledge of the shift characteristics is available. Knowing that the shift is systematic (as in the case of the GC example) the alignment procedure can benefit from a linear shift correction before the actual COW alignment. This method moves the intact chromatogram a number of data points to the left or right (flexibility) and the correlation to the reference sample is calculated [13].

The peak shape plays a major role in the alignment procedure. Overloading the stationary phase in a chromatographic system can cause peaks to tail or front in such a way that the peak shape changes significantly, whereas the area is kept proportional to the concentration of the eluting compound. Aligning peaks of different shapes (sample towards reference chromatogram) can cause a dramatic change in the area of peaks of the sample chromatogram, as the interpolation of data points will be guided by the shape of

DOI: 10.1002/cem

peaks in the reference chromatogram. Thus, the COW algorithm is mostly suitable for chromatographic data containing peaks that do not change shape across chromatograms caused by, for example, severe peak tailing.

Another major issue is the number of data points per peak. If this number is small (i.e. 7–10 data points) for certain peaks in the chromatogram the interpolation steps can result in significant changes in area and thus increase the uncertainty of the area estimation even for replicated samples. This can be remedied by interpolating to a higher resolution before the alignment procedure. However, increasing the number of data points obtained during the chromatographic run has a cost—the alignment process becomes more time-consuming.

The warping effect value is the sum of the simplicity and peak factor values. A high value does not imply both optimal alignment and area preservation, but rather an optimal compromise. This was depicted in Tables 1–3 where other combinations than the optimal one found from the optimisation routine could hold a higher simplicity or peak factor value. Thus, dependent on the purpose of the alignment of the chromatographic profiles, the focus in the combined expression (warping effect) routine could be put on either of these two quantitative terms by individual weighting.

In this paper, we show that automatic parameter selection for the COW alignment correction can lead to good results for chromatographic data. All calculations shown have been performed in MATLAB® version 7/R2006a (The MathWorks, Inc., Natick, Ma, USA) using a Pentium 4, 3.40 GHz processor with 2 GB of RAM. The simplicity and optimisation routines can be freely downloaded from Reference [13].

Acknowledgements

The authors acknowledge Merete Møller Nielsen for providing the HPLC-Fluorescence data. Thomas Skov received financial support from ARLA Foods Amba and The Royal Veterinary and Agricultural University under the Graduate School of Food Science and Technology, LMC.

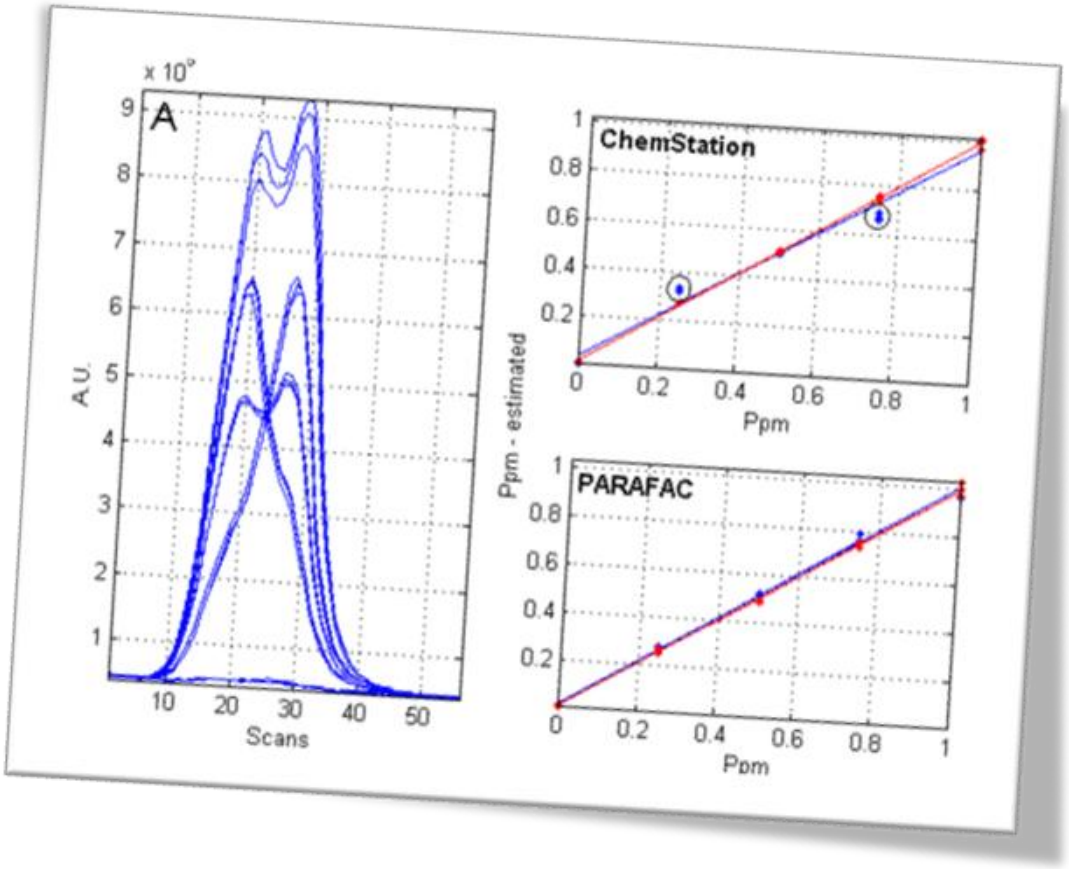
REFERENCES

1. Nielsen NPV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998; **805**: 17–35.
2. Bylund D, Danielsson R, Malmquist G, Markides KE. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 2002; **961**: 237–244.
3. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.* 2004; **18**: 231–241.
4. Pravdova V, Walczak B, Massart DL. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* 2002; **456**: 77–92.
5. Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal. Chem.* 1987; **59**: 649–654.
6. Grung B, Kvalheim OM. Retention time shift adjustments of two-way chromatograms using Bessel's inequality. *Anal. Chim. Acta* 1995; **304**: 57–66.
7. Wong JWH, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.* 2005; **17**: 5655–5661.
8. Tomasi G. *Practical and computational aspects in chemometric data analysis. Doctoral Thesis, The Royal Veterinary and Agricultural University (KVL)* 2006, http://www.models.life.ku.dk
9. Henrion R, Andersson CA. A new criterion for simple-structure core transformations in N-way principal components analysis. *Chemom. Intell. Lab. Syst.* 1999; **47**: 189–204.
10. Christensen J, Tomasi G, Hansen AB. Chemical fingerprinting of petroleum biomarkers using time warping and PCA. *Environ. Sci. Technol.* 2005; **39**: 255–260.
11. Johnson KJ, Prazen BJ, Young DC, Synovec RE. Quantification of naphthalenes in jet fuel with GC × GC/Tri-PLS and windowed rank minimization retention time alignment. *J. Sep. Sci.* 2004; **27**: 410–416.
12. Spendley W, Hext GR, Himsworth FR. Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation. *Technometrics* 1962; **4**: 441–461.
13. *The Quality & Technology Website: http://www.models.kvl.dk [August 2006].*

DOI: 10.1002/cem

PAPER II

Skov, T. and Bro, R. **2008**. Solving fundamental problems in chromatographic analysis. *Analytical and Bioanalytical Chemistry*, 390 (1): 281-285.



Anal Bioanal Chem (2008) 390:281–285
DOI 10.1007/s00216-007-1618-z

TRENDS

Solving fundamental problems in chromatographic analysis

Thomas Skov · Rasmus Bro

Published online: 24 October 2007

Introduction

Chromatographic analytical systems are being increasingly used for analysis of complex samples, for example, in metabonomics and food analysis. Hyphenated separation techniques (multidimensional) such as LC–MS, GC–MS, and HPLC–UV are intensively used for obtaining detailed qualitative and quantitative information. The samples are typically of a much more complex nature than in traditional analytical chemical applications and this poses problems for the traditional approaches for handling such data.

Multidimensional techniques have been used for many decades and many approaches have been put forward to deal with data when having perfect resolution of the eluting peaks. However, with more complex samples and/or the need for faster chromatographic runs, perfect separation cannot always be achieved (Fig. 1). Traditional data analysis relying on resolved peaks would, even with additional mass spectral information, fail to find the underlying analytes if the overlap is too severe. Also, analytes having the same dominating fragments in the mass spectrum would be a problem even with less severe overlap of the peaks. It should be noted that more advanced analytical techniques such as tandem MS–MS or high resolution MS can be used to some degree to resolve overlapping peaks even for low mass and closely related compounds.

T. Skov (✉) · R. Bro
Quality and Technology, University of Copenhagen,
Rolighedsvej 30,
Frederiksberg C
1958 Denmark
e-mail: thsk@life.ku.dk

Other fundamental problems such as low signal-to-noise ratio of peaks resulting in background interference in the mass spectrum, varying baseline due to column bleed and peak shifts due to degrading columns can be hard to handle with traditional data-analysis tools.

This paper deals with ways of solving these fundamental problems showing how more advanced data analytical methods can be a good (and often better) alternative to traditional tools such as those offered in commercial chromatographic software (e.g. ChemStation). Focus here will be on resolving the data from gas chromatographic systems into chemically meaningful estimates, whereas alternative more exploratory fingerprinting will not be explicitly covered. The methods presented are also valid for other chromatographic techniques such as LC and HPLC.

Analysis of chromatographic data

The first step in solving these fundamental problems is to transfer the data to a numerical computing environment such as MATLAB. This data transfer is actually a problem in itself. The most common chromatographic format is a so-called netCDF format, which is a format that most manufacturers support. However, the transfer to other software is not straightforward and requires advanced toolboxes and often basic knowledge in programming [1]. A recent toolbox for data analysis of metabolome data [2] handles netCDF data in a black box environment and the often cited netCDF toolbox for MATLAB version 6 [3] uses rather advanced features. None of these solutions, however, are accessible to laymen. In order to stimulate research in advanced chemometric data analysis, a free and documented toolbox has been developed, which makes the

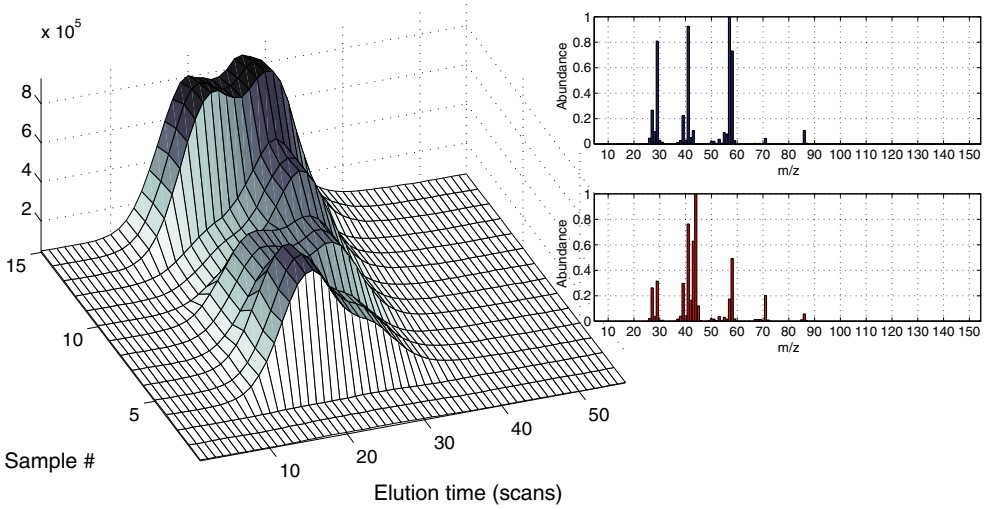


Fig. 1 *Left*: Example of total-ion-current chromatograms (TIC) for a mixture of two important aroma compounds found in cheese (*A*: 2-methylbutanal and *B*: 3-methylbutanal). Here analyzed as a mixture in heptane on a polar 30 m long DB-Wax capillary column with an internal diameter of 0.25 mm, and with a 0.25 μm film (thickness). Samples 1–3: 0 ppm of both, samples 4–6: 0.75 ppm of *A* and 0.25 ppm of *B*, samples 7–9: 0.50 ppm of both, samples 10–12: 0.25 ppm of *A* and 0.75 ppm of *B*, and samples 13–15: 1.00 ppm of both. *Right*: EI Mass spectra of the two aroma compounds—*top*: 2-methylbutanal and *bottom*: 3-methylbutanal

import of data a simple operation. This function called iCDF (import CDF) is available from [4].

Preprocessing of the data

Before the actual data analysis, fundamental problems such as varying baseline and peak shifts must be handled. Several attempts have been made to correct for the effects of column bleed; one of the most cited and used approaches was described by Eilers [5]. This method, which has been shown to be very effective for chromatographic data, uses asymmetric least-squares smoothing, by penalizing regions with signal in the chromatogram, to find the data points corresponding to the baseline. This approach has been extended to two-dimensional data with success for 2D gel electrophoresis [6]. Methods such as this remove baseline from the chromatograms so that ideally only contributions from the eluting analytes are contained in the signals. In commercial software packages the baseline is often approximated by a straight line from the peak start to peak end and this can lead to biased results if peaks are eluting on a significant temperature ramp or if the peak is difficult to detect due to noise.

Alignment (or warping) of shifted peaks can be performed in various ways. In commercial software packages, a window around the peak is often used to search for identical (or similar) mass spectra among samples in order to assign the peaks correctly. However, this approach is highly affected by the degree of shift and thus, a way to align the peaks prior to this peak detection approach would be preferred. One effective method for chromatographic data is the piecewise alignment method correlation optimized warping (COW), which can handle non-systematic global shifts [7, 8]. With the possibility of finding appropriate parameters for the warping algorithm in a semi-automated way, this preprocessing tool is easily applied even for the less experienced user [9]. For commercial GC–MS instruments using a quadrupole mass analyzer, the warping parameters can be found from the TIC chromatogram and applied to each individual ion chromatogram, as it is assumed that no shift in the mass spectral dimension takes place. This can speed up analysis significantly. No commercial toolbox has been presented to transfer aligned data back to the GC–MS software, which could be helpful for traditional users of GC–MS software. However, for one-dimensional chromatographic data LineUp from Infometrix can both read and write in formats supported by GC software [10].

Peak detection

It is often necessary to detect the peaks in chromatograms and to estimate the number of overlapping peaks in a given window. Several peak-detection methods using information from derivatives of elution profiles have been put forward and found to work well for chromatographic data [11, 12]. Other approaches using the information from the mass spectral dimension have been proposed to either detect peaks or to match peaks from different chromatograms if no alignment has been carried out [12, 13]. The window size should not necessarily be of the same size throughout the chromatograms, and the user has to ensure that no peaks are split between adjacent windows.

A comprehensive and quite detailed approach to automating the steps discussed so far can be found elsewhere [12, 13].

Multiway data needs multiway methods

Multidimensional chromatographic measurements provides data of more than the traditional two dimensions (samples × variables) and this gives new possibilities with regard to the information that can be extracted. There are methods that make specific use of the so-called second-order or multiway structure of such data, a feature that can be used to quantify analytes in the presence of unknown interfering chemical compounds [14] which would otherwise require more comprehensive calibration. Especially parallel factor analysis (PARAFAC) [15] has been used to model multiway chromatographic data to get both qualitative and quantitative information [16–19] even when several peaks are overlapped. PARAFAC is a natural extension of principal component analysis (PCA). In PCA underlying features are found using orthogonality constraints. In such analysis true chemical information cannot be expected in the components because of what is known as the rotational ambiguity, where the same fit of the model can be achieved if rotating the model parameters. With PARAFAC this is not the case and true and unique estimates of single-analyte chromatograms and spectra can be found from overlapped peaks provided the data follow the trilinear structure of the model. This is shown in Fig. 2 where the pure analyte spectra and elution profiles are found directly using no information other than the experimental measurements.

In Fig. 2 it can be seen that PARAFAC provides pure chromatograms and spectra for the two eluting analytes (and the background). When this is the case, the scores (the amount of the underlying features) are estimates of the relative concentrations. Hence, both qualitative and quantitative information is obtained directly from the mixtures. This has also been termed "mathematical chromatography". An interesting feature can be observed in Fig. 2. Instead of removing the baseline, which can be rather tricky for multidimensional chromatographic data, PARAFAC simply models this as an individual factor in a much more rational way than ordinary subtraction of a fitted baseline.

One aspect that the PARAFAC model does not include is the absolute scale of the scores and thus it is still important to have an idea of the absolute concentration of the analytes (e.g. the concentration of 3-methylbutanal in one of the samples in Fig. 1). Figure 3 shows the estimated concentration based on PARAFAC modeling and based on a traditional ChemStation analysis of the peaks illustrated in Fig. 1 and from two other cases with severe peak overlap of important aroma compounds in cheese. In addition, the results from so-called PARAFAC2 are also shown for Fig. 3b because it allows direct handling of shifts and peak changes without prior alignment.

In Fig. 3 the traditional method performs quite well but it will underestimate or overestimate the concentration when peaks are overlapped (especially for a shoulder peak of lower concentration) and a difference in the concentrations is observed (Figs. 3a and b) [20, 21].

Figure 3b illustrates that even the advanced method can have problems. In this situation shift in peak maxima and

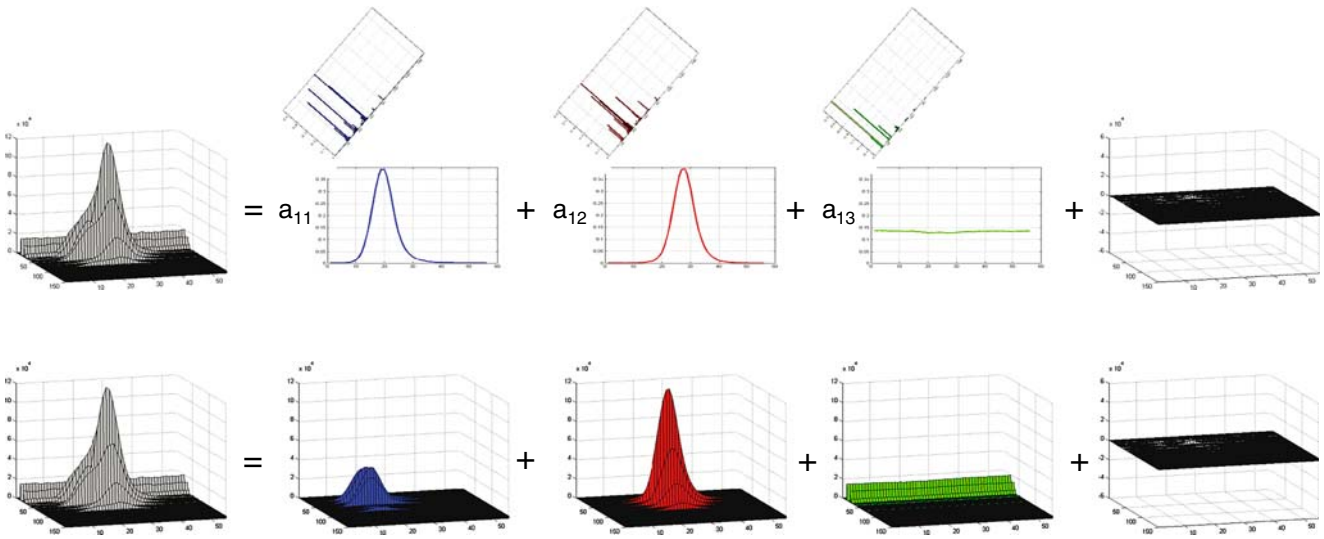


Fig. 2 PARAFAC decomposition (using three factors) of the mixture of the two aroma compounds shown in Fig. 1. For visualization only sample eleven (*A*: 0.25 ppm and *B*: 0.75 ppm) is shown here. *Top*: PARAFAC model shown with loading matrices for the pure chromatographic and mass spectral profiles for the two analytes and baseline found in sample eleven. *Bottom*: PARAFAC model shown with multiplied loadings. The unexplained part, the residuals, is shown to the *outer right*. Note that the magnitude of the ordinate axes for the three factors (multiplied loadings) and the residuals are the same to indicate the model performance

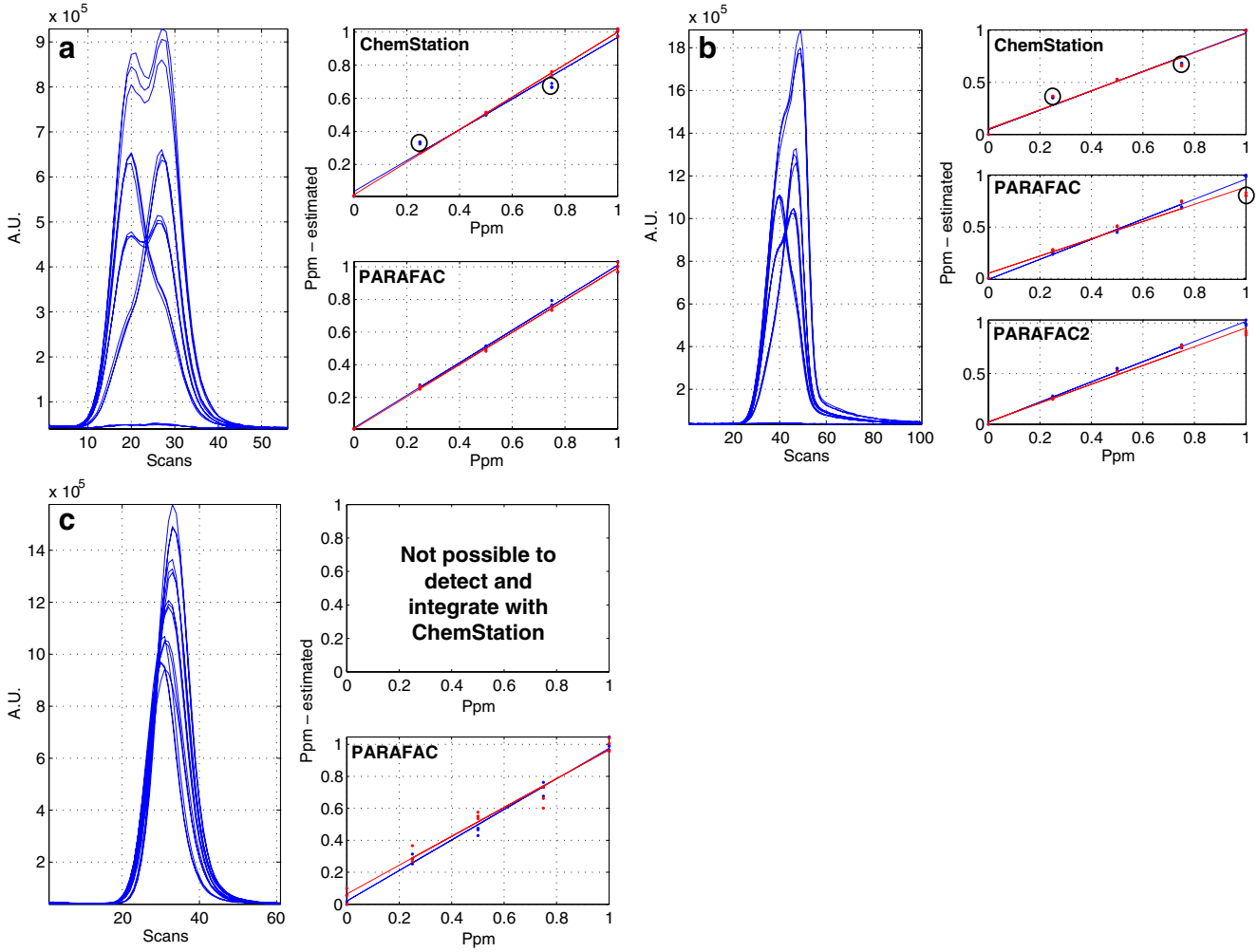


Fig. 3 Calibration models for the estimated peak areas using Integration (ChemStation) or score values (PARAFAC/PARAFAC2). In **a**–**c** the *left plot* shows the TICs of the three-way array for the peaks analyzed. **a**: See Fig. 1 for details. *Top*: ChemStation, *Bottom*: PARAFAC. **b**: 3-methylbutanol and 2-methylbutanol: *Top*: ChemStation, *Middle*: PARAFAC, *Bottom*: PARAFAC2. **c**: Diacetyl and 2-pentanone. *Top*: ChemStation, *Bottom*: PARAFAC. The *circles* highlight the differences between the two approaches, as explained in the text

peak broadening mainly related to the change in concentration of the samples were observed. This caused a deviation from trilinearity resulting in lower score values than appropriate. This problem can be solved using the more flexible PARAFAC2 model. Unlike PARAFAC, the PARAFAC2 model can model the elution profiles in different experiments individually and hence handle possible differences in their shape [22, 23]. As seen from Fig. 3b the PARAFAC2 model solved the problem with deviation from trilinearity very well, providing a significantly improved calibration curve.

In situations with very similar mass spectra (same dominating fragments) and severe overlap, the traditional method would fail to find the two analytes whereas PARAFAC performs well as long as differences in the ratios of the fragments are present (Fig. 3c).

Conclusion

The major advantages of solving fundamental problems in chromatographic data outside commercial GC software are better chromatographic understanding (pure chromatograms and spectra are readily assessable even with severe shifted data), more flexibility (more methods to try out), and the possibility of automating the individual steps and through the latter to keep it simple (all preprocessing steps and model parameters can be evaluated from a chromatographic point of view) for the user that might not be familiar with the advanced methods.

PARAFAC coupled with proper data preprocessing will provide an estimate of the relative concentrations of each eluting analyte and the spectrum and elution profile of each analyte.

However, some consideration is still needed before PARAFAC can be applied with success. Overlapping peak

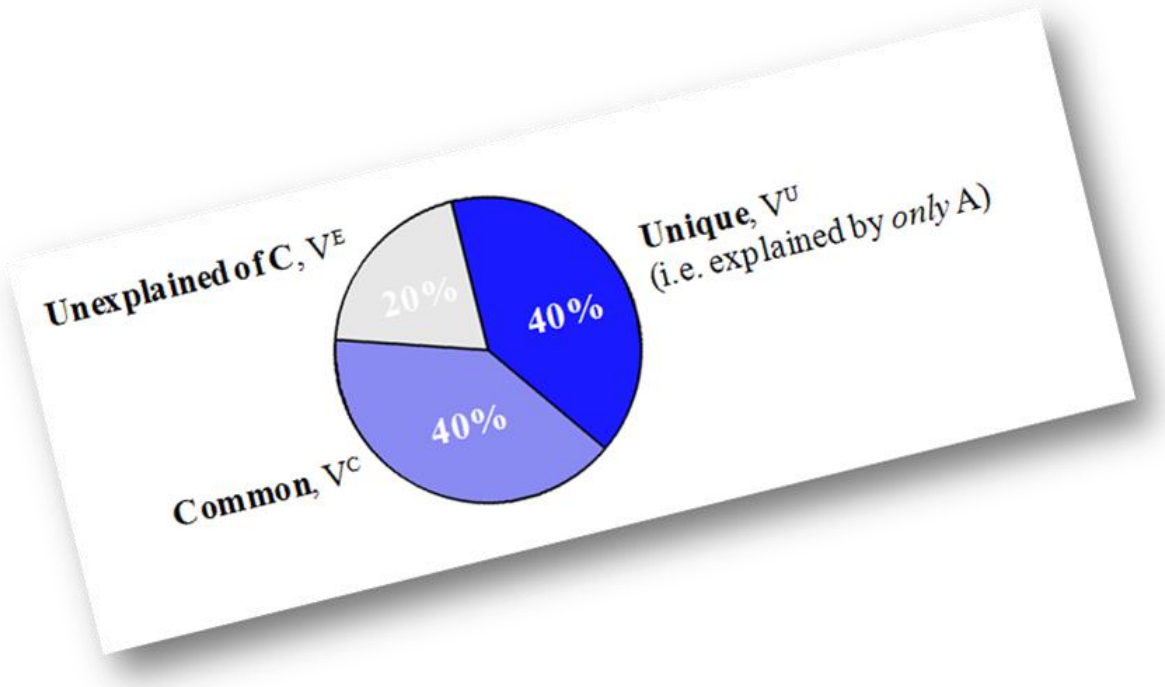
regions or changes in peak shape due to overload of the column can make the alignment procedure less efficient and thereby disturb the trilinear structure of the data. These fundamental problems are complex to solve using the preprocessing methods of today, but can be handled by the more advanced and flexible PARAFAC2.

References

1. Rew R, Davis G (1990) IEEE Comput Graphics Appl 10:76–82. Unidata [July 2007, URL: http://www.unidata.ucar.edu/software/netcdf/]
2. Bunk B, Kucklick M, Jonas R, Munch R, Schobert M, Jahn D, Hiller K (2006) Bioinformatics 22:2962–2965
3. US Geological Survey (USGS), Woods Hole, NetCDF toolbox for MATLAB 6, [July 2007, URL: http://mexcdf.sourceforge.net/index.html]
4. iCDF, Quality & Technology, Department of Food Science, University of Copenhagen, Denmark [September 2007, URL: http://www.models.life.ku.dk/source/iCDF]
5. Eilers PHC (2004) Anal Chem 76:404–411
6. Kaczmarek K, Walczak B, de Jong S, Vandeginste BGM (2005) Acta Chromatogr 15:82–96
7. Nielsen NPV, Carstensen JM, Smedsgaard J (1998) J Chromatogr A 805:17–35
8. Tomasi G, van den Berg F, Andersson C (2004) J Chemom 18:231–241
9. Skov T, van den Berg F, Tomasi G, Bro R (2006) J Chemom 20:484–497
10. LineUp, Infometrix, [July 2007, URL: http://www.infometrix.com/software/LU2Specs.pdf]
11. Vivo-Truyols G, Torres-Lapasio JR, van Nederkassel AM, Heyden YV, Massart DL (2005) J Chromatogr A 1096:133–145
12. Dixon SJ, Brereton RG, Soini HA, Novotny MV, Penn DJ (2006) J Chemom 20:325–340
13. Stein SE (1999) J Am Soc Mass Spectrom 10:770–781
14. Boque R, Ferre J (2004) LC–GC Eur 17:402–407
15. Bro R (1997) Chemom Intell Lab Syst 38:149–171
16. Bylund D, Danielsson R, Malmquist G, Markides KE (2002) J Chromatogr A 961:237–244
17. Hoggard JC, Synovec RE (2007) Anal Chem 79:1611–1619
18. Bro R (2006) Crit Rev Anal Chem 36:279–293
19. Johnson KJ, Rose-Pehrsson SL, Morris RE (2004) Energy Fuels 18:844–850
20. Bicking MKL (2006) LC–GC North Am 24:402–414
21. Bicking MKL (2006) LC–GC North Am 24:604–616
22. Kiers HAL, Ten Berge JMF, Bro R (1999) J Chemom 13:275–294
23. Bro R, Andersson CA, Kiers HAL (1999) J Chemom 13:295–309

PAPER III

Skov, T., Ballabio, D., and Bro, R. **2008**. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Analytica Chimica Acta*, 615 (1): 18-29.



ANALYTICA CHIMICA ACTA 615 (2008) 18–29

available at www.sciencedirect.com

ScienceDirect

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

Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks

Thomas Skov[a,*], *Davide Ballabio*[b], *Rasmus Bro*[a]

[a] *Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[b] *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1-20126 Milano, Italy*

ARTICLE INFO

Article history:
Received 28 January 2008
Received in revised form
17 March 2008
Accepted 20 March 2008
Published on line 29 March 2008

Keywords:
Multiblock
Partial least squares
Unique variation

ABSTRACT

More than one multi-informative analytical technique is often applied when describing the condition of a set of samples. Often a part of the information found in these data blocks is redundant and can be extracted from more blocks. This study puts forward a method (multiblock variance partitioning—MVP) to compare the information/variation in different data blocks using simple quantitative measures. These measures are the unique part of the variation only found in one data block and the common part that can be found in more data blocks. These different parts are found using PLS models between predictor blocks and a common response. MVP provides a different view on the information in different blocks than normal multiblock analysis. It will be shown that this has many applications in very diverse fields such as process control, assessor performance in sensory analysis, efficiency of preprocessing methods and as complementary information to an interval PLS analysis. Here the ideas of the MVP approach are presented in detail using a study of red wines from different regions measured with GC–MS and FT-IR instruments providing different kinds of data representations.

1. Introduction

The development of novel analytical techniques to analyze a wide range of complex and diverse samples has progressed significantly in the last decade. Today all characteristics such as smell, taste, appearance, structure, texture and micro flora of a food product can be measured using multiple measurement techniques (spectroscopy, chromatography, etc.) based on various chemical and physical principles. Often several methods are used to give a detailed and complete description of the products investigated. The information in the data from different instruments are frequently regarded as being independent—i.e. to describe different phenomena in the sample, thus adding supplementary information each time a new instrument is used. Multivariate data from these analytical techniques are often analyzed with suitable bilinear chemometric methods such as principal component analysis (PCA) [1]. This is done in order to evaluate and find classes, outliers, correlations with other signals, etc. The data can be linked as a predictor block to responses (e.g. NIR spectral measurements linked to the response variable water content) and then analyzed by means of regression methods, such as partial least squares (PLS) [2].

With a comprehensive product description, using many analytical techniques on the same set of samples, multiple predictor blocks as well as multiple response blocks are pro-

* *Corresponding author*. Tel.: +45 35 33 37 39.
E-mail address: thsk@life.ku.dk (T. Skov).
0003-2670/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2008.03.045

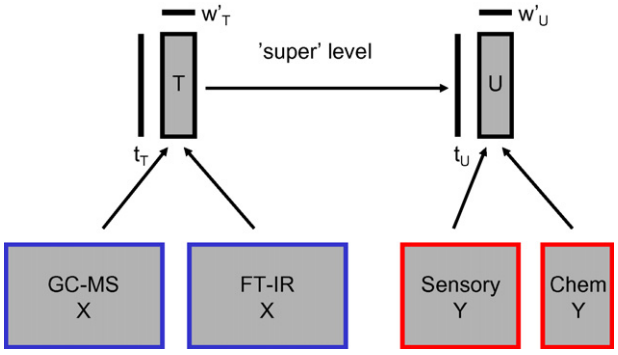


Fig. 1 – Conceptual idea behind multiblock PLS—here with two predictor blocks and two response blocks. The X-blocks are linked to Y-blocks through scores and loadings weights on the block level which are then suitably combined on the super-level through a super level model.

duced. The easiest way to handle several predictor blocks is to augment the different blocks into one single predictor matrix and use PCA or PLS on that. This approach is simple and straightforward and works well in many situations. However, merging several blocks of different types, i.e. discrete and continuous signals, can be problematic: a large number of variables can be difficult to interpret, and the scaling of individual blocks consisting of different variables can have a huge impact on the final results. One way to circumvent these issues is to treat each predictor block independently in PCA or PLS.

Several multiblock methods have been described in the literature and especially in PLS regression (see Fig. 1) the use of such techniques has shown to be very successful [3–9]. Tenenhaus and Vinzi [5] reviewed the use of multiblock analysis, focusing on multiblock methods where a specific criterion (i.e. covariance, correlation) is optimized. Westerhuis et al. [7] described multiblock PCA and PLS from an algorithmic point of view and exemplified the findings with several case studies. The PLS approach uses a so-called super score deflation technique where an initial guess of super scores, i.e. best descriptive scores for all **X**-blocks, are used in deflating the **X** [7] or the **Y** matrices [8]. This is done iteratively until convergence. The latter approach with deflation of the **Y**-block has been shown to be the most appropriate in multiblock PLS models, providing the same predictions as a standard PLS model with one augmented **X**-block (using proper scaling!). For a detailed description of the issue in deflating of either **X** or **Y** see [7,8].

As stated previously, prediction of a specific response using multiblock PLS provides results similar to a conventional PLS model. However, multiblock PLS gives superior results compared to single block PLS with respect to interpretability because it is possible to 'zoom in' on separate blocks using the block scores calculated for each **X** matrix and evaluate the specific relation to the **Y**-block in presence of the other **X** matrices.

The multiblock PLS models do not provide explicit quantitative measures about redundant information from the multiple **X**-blocks. Stated differently, they do not describe how much common variation (joint, overlapping, correlating and redundant) and unique variation can be found in the predictor blocks. These measures can be very advantageous when several blocks are available to see, e.g. if one or more blocks are sufficient or all are needed to describe the responses.

Some designated chemometric techniques have been put forward to find joint and unique variation from several data blocks. O2PLS (orthogonal PLS) is a modification of the OPLS method that provides separate models for both joint and unique (non-correlated/orthogonal) variations between two blocks of data with more than one variable in **Y** [10]. However, this method suffers from the before-mentioned scaling issues as all **X**-blocks must be augmented prior to the prediction of **Y** if the unique part of more than one **X**-block is to be evaluated. Another new multiblock related method uses the scores from two PLS2 models when predicting a **Y**-block followed by comparing the scores from these models to find the best correlation by iteratively rotating the individual models until convergence [11]. However, this method focuses on the prediction of **Y** and is not capable of dealing with more than two **X**-blocks. A multiblock method called serial PLS (S-PLS) presented by Berglund and Wold [3] and further investigated by Felicio et al. [4] evaluate the common and unique part of two **X**-blocks predicting the same **Y**-block. S-PLS regresses the residuals of **Y** from a model between the first **X**-block and **Y** using the second as predictor and vice versa until the best prediction of **Y** has been found. This has been shown to give superior results compared to ordinary classical PLS [4] without any scaling issues, but also that the improvements are small, the algorithm rather slow and no suggestion has been put forward to deal with more than two **X**-blocks.

In this paper we will not be focusing on the predictive aspects of multiblock PLS but we will use a new variant of multiblock PLS to help establish quantitative measures for the redundancy in information between several **X**-blocks. A new approach termed multiblock variance partitioning (MVP) for analyzing and comparing the information found in the several **X**-blocks is introduced. Data from GC–MS and FT-IR measured on 44 red wine samples made from 100% Cabernet Sauvignon grapes harvested in three different geographical regions (South Africa, South America and Australia) will be studied. Several studies deal with the characterization of wines using methods related to both taste and aroma either individually or as complementary techniques [12–17]. Although highly specific information can be found from both techniques (e.g. aroma compounds and alcohol concentration) that characterize the individual red wines, the approach here will be to use data as fingerprints of the wine and to quantify how the data blocks supplement each other in providing information on the measured wine samples.

Data are evaluated using partial least squares (PLS) [2] models between one data block, selected as **Y** (the response), and each of the remaining blocks as **X** (predictor block) by evaluating how much variation in **Y** can be predicted from the remaining data blocks. This evaluation is done focusing on one of the **X**-blocks (in turn) describing to what degree this matrix can describe **Y** and to what extent the remaining data can add to the description. As a result, for each data block, the variation in the original **Y**-block can be split into three parts: (1) the *unexplained* part not described by the additional predictor data blocks, (2) the common part (*common* variation) that

Table 1 – Geographical origin of the analyzed red wines

Origin	Wine samples
Argentina	6
Chile	15
Australia	12
South Africa	11
Total	44

holds the variation found in the other **X**-blocks as well, (3) and the *unique* part that can only be found in the specific predictor block under consideration.

Our approach of using individual PLS models is selected to avoid the block-scaling issue when dealing with multiblock data or when augmenting several **X**-blocks. The scaling-of-the-blocks issue has been discussed in many papers where the observation is made that very different results and interpretations are found depending on the scaling method of choice for block-scaling dependent methods [3–11,18].

2. Materials and methods

2.1. Wine samples

Red wines, 44 samples, produced from the same grape (100% Cabernet Sauvignon to eliminate the influence of different grape varieties), harvested in different geographical areas, have been collected from local supermarkets in the area of Copenhagen, Denmark. Details on the geographical origins and number of wine samples analyzed are given in Table 1.

2.2. Instrumentation

2.2.1. GC–MS

Dynamic headspace gas chromatography mass spectrometry (HS-GC–MS) was used to measure the aroma profile of the wine samples.

2.2.1.1. Sample preparation. 10 mL of each wine sample, without further sample preparation, were added directly into a 100 mL purge flask and 2 mL 4-methyl-1-penthanol in a water solution (50 mg L^{-1}) was added as internal standard. The samples were equilibrated to 30 ± 1 °C in a circulating water bath and then purged with nitrogen (75 mL min^{-1}) for 20 min. The volatile compounds were collected on a Tenax-TA trap. The trap contained 250 mg of Tenax-TA with mesh size of 60/80.

2.2.1.2. Desorption of volatile compounds. The trapped volatiles were desorbed using an automatic thermal desorption unit (ATD 400, PerkinElmer, Norwalk, USA). Primary desorption was carried out by heating the trap to 250 °C with a flow (60 mL min^{-1}) of carrier gas (He) for 15.0 min. The stripped volatiles were trapped in a Tenax TA cold trap (5 °C), which was subsequently heated to 300 °C (secondary desorption). This allowed for rapid transfer of volatiles to a gas chromatograph–mass spectrometer through a heated (225 °C) transfer line.

2.2.1.3. GC conditions. Separation of aroma compounds was carried out on gas chromatography system (HP 6890 GC) with a 30 m long DB-Wax capillary column with an internal diameter of 0.25 mm and a 0.25 μm film thickness. The column flow rate was 1.0 mL min^{-1} using helium as a mobile phase. The column temperature program was: 10 min at 45 °C, from 45 to 240 °C at 6 °C min^{-1}, and 10 min at 240 °C. An injection split ratio of 1:50 was used for all the experiments.

2.2.1.4. MS conditions. The GC was equipped with a mass spectrometric detector (Agilent 5973 Mass Selective Detector) operating in the electron ionization (EI) mode at 70 eV. Mass-to-charge ratios between 15 and 300 were scanned twice for each GC scan. GC inlet was held at 250 °C and the MS transfer line maintained at a temperature of 280 °C.

In both the sample preparation and GC run the samples were randomized to minimize the introduction of systematic effects in data. Duplicates of all samples were carried out and the average used for the data analysis.

GC–MS data were initially explored and analyzed in Agilent ChemStation software (Enhanced ChemStation G1701DA Version D.00.00.38, Agilent Technologies Inc., Palo Alto, CA, USA) to identify aroma compounds (using Wiley275.L, HP product no. G1035A) and to integrate the corresponding peak areas. Subsequently, data were exported as netCDF files (i.e. AIA format in ChemStation), and using *i*CDF [19], a tool to simplify the import of netCDF files, imported into MATLAB® version 7.3 (R2006b) (The MathWorks Inc., Natick, Ma, USA), where processing with advanced data analyses was conducted (baseline correction, peak alignment, modelling, classification, etc.). The *i*CDF and peak alignment tools used can be downloaded for free at www.models.life.ku.dk and more information can found in [19,20].

2.2.2. WineScan

2.2.2.1. Sample preparation. Approximately 100 mL of wine was transferred to small 100 mL bottles and analyzed directly by placing them in the sample holder of a WineScan instrument (WineScan™ FT120 Basic, FOSS Analytical). No sample preparation is necessary, as the WineScan instrument is build to analyze the finished wine directly from the bottle.

2.2.2.2. WineScan conditions. A WineScan FT120 instrument (Foss Electric, Hillerød, Denmark) that employs a Michelson interferometer was used to generate the FT-IR spectra. Samples (7 mL) were pumped through the CaF_2-lined cuvette (optical path length 37 μm), which is placed in the heater unit of the instrument. The temperature of the samples is brought to exactly 40 °C before analysis. Samples were scanned from 5011 to 929 cm^{-1} with 4 cm^{-1} intervals (i.e. 1056 data points per spectrum), which includes a small section of the near-IR region.

The intensity of the IR beam transmitted through a sample was recorded at the detector and used to generate an interferogram that is calculated from a total of 20 scans before being processed by Fourier transformation to generate a single beam transmittance spectrum. Background absorbance in the wine sample (which includes the absorbance of water) is corrected through the use of a Foss Zero Liquid S-6060, which is scanned prior to the wine sample. The single beam

Table 2 – Quality parameters measured on the WineScan instrument and used in MVP (units shown in brackets)

#	Quality parameter
1	Ethanol (vol.%)
2	Total acid ($g\,L^{-1}$)
3	Volatile acid ($g\,L^{-1}$)
4	Malic acid ($g\,L^{-1}$)
5	pH
6	Lactic acid ($g\,L^{-1}$)
7	Rest sugar (glucose + fructose) ($g\,L^{-1}$)
8	Citric acid ($mg\,L^{-1}$)
9	CO_2 ($g\,L^{-1}$)
10	Density ($g\,mL^{-1}$)
11	Total polyphenol index
12	Glycerol ($g\,L^{-1}$)
13	Methanol (vol.%)
14	Tartaric acid ($g\,L^{-1}$)

transmittance spectrum of the zero liquid is stored on the computer of the instrument, and the ratio of the sample spectrum to the zero liquid spectrum at each recorded data point is used to generate the final transmittance spectrum. The transmittance spectra were converted into absorbance spectra where the water absorption bands in the regions between 1545–1710 cm^{-1} and 2968–3620 cm^{-1} were excluded.

2.2.2.3. Calibration models and predicted quality parameters. WineScan comes with a build-in calibration model capable of predicting the concentration of the listed quality parameters in a specific concentration range. The calibration models (i.e. which regions in the FT-IR spectrum are used to predict the individual parameters) are not available to the user, and only the predicted values can be extracted. The 14 calibration models are made from a large database of commercially available red wines and should thus meet the requirements set by an experiment evaluating wines bought in the supermarket. For further information on the WineScan the reader is referred to the FOSS Analytical webpage: http://www.foss.dk/.

Duplicates of all samples were carried out and the average used for the data analysis. For all wine samples 14 quality parameters were predicted using the FOSS *WineScan* build-in calibration models (Table 2).

2.3. Data structures

The data from both GC–MS and FT-IR measurements are complex and highly multivariate. As the first step in simplifying and understanding the data, different condensed representations of the data are made. The GC–MS data are represented as either (1) integrated peak areas, (2) GC time-elution profiles summed over the mass dimension, or (3) mass profiles summed over the elution time dimension. These three different representations will all be investigated as descriptor **X**-blocks. Likewise, the IR data will be represented as either (4) predicted quality parameters or as (5) raw spectra (Table 2) produced from the instrument calibration models. It is reasonable to prefer the simpler representations (peak areas for GC–MS and quality parameters for IR) as these represent the most direct chemical information level. However, using these representations is only valid to the extent that they carry sufficient information. Whether this is valid and sufficient or not is a part of what the new multiblock variance partitioning approach can help establishing.

2.3.1. GC–MS data blocks

For each sample a mass spectrum scan (*m*/z: 5-204) measured at 2700 elution time-points was obtained providing a data cube of size $44 \times 2700 \times 200$. Condensing the data by summing over the mass dimension is a generic approach when analyzing GC–MS data with bilinear models. For samples of low complexity, like the ones studied here, it can be assumed that there is no loss of information in data. On the other hand, summing over the elution time dimension, some loss in information is expected due to the high degree of fragmentation of the sample fractions (i.e. eluting analytes) entering the mass detection system (different analytes will provide similar ion fragments). However, with all three GC data representations included it can be investigated how much information is maintained from the different data condensations.

In the ChemStation software 57 aroma compounds were identified and used as peak areas for the fingerprint. Instead of using ChemStation more advanced methods could also have been used to extract the peak areas directly from the three-way data [19], however, this approach was not pursued in this study. Table 3 shows the dimensions of the three GC related data blocks.

2.3.2. IR data blocks

All 14 quality parameters found from the IR spectra measured on the WineScan were selected and used for the data comparison. The IR data blocks are also specified in Table 3.

2.4. Data presentation

In Figs. 2–6 one red wine sample is depicted using the two measurement techniques and data structures presented in Table 3. In Fig. 3 a typical chromatogram (total ion count chromatogram—TIC) of a wine sample is shown with the corresponding peak areas of the aroma compounds extracted from the ChemStation software (the identification of the aroma compounds is not included here).

GC–MS and FT-IR data can be downloaded from www.models.life.ku.dk/research/data.

2.5. Multiblock variance partitioning

2.5.1. Notation

Scalars are indicated with lower-case italics (e.g. x_{ijk}) and vectors with bold lower-case characters (e.g. **x**). Matrices are denoted **X**, **Z**, **Y** or **Q** ($I \times J$), where I is the number of objects, J the number of variables.

In particular:

- **Y** is the response matrix/vector (one of the data blocks currently under study),
- **X** is the **X** matrix used to primarily predict **Y**,
- $\mathbf{Z}_k$ is the kth additional data blocks ($k = 1, \ldots, K$) that is considered in comparison to the unique part of **X**.

Then,

- $\mathbf{Y}_U$ is the predicted matrix/vector using **X**,

Table 3 – Dimensions (samples × variables), source, and type of the 5 data blocks used

#	Data block	Source	Type of data	Dimensions	Preprocessing
1	GC peak areas	GC	Discrete	44 × 57	Auto scaled
2	GC elution profiles		Continuous	44 × 2700	Centered
3	GC mass profiles		Continuous	44 × 200	Centered
4	IR quality parameters	IR	Discrete	44 × 14	Auto scaled
5	IR spectra		Continuous	44 × 842	Centered

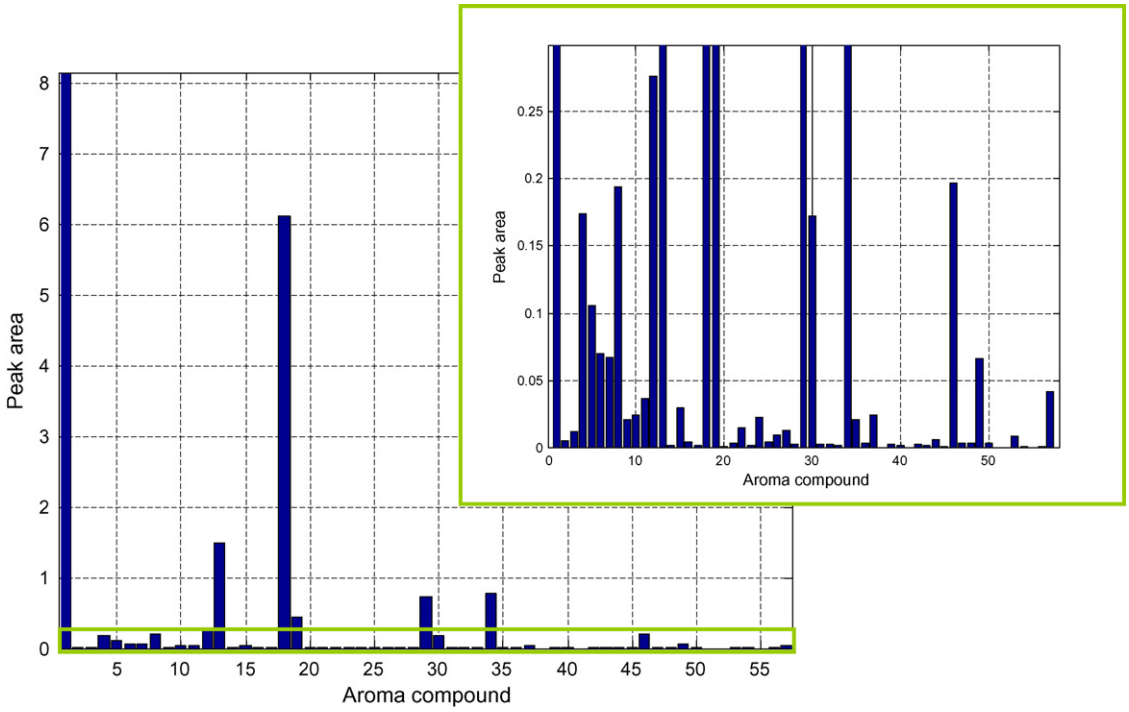


Fig. 2 – Data block #1: GC peak areas. Relative concentration of aroma compounds found in one red wine sample. Top right: Zoom-in on peaks with small areas.

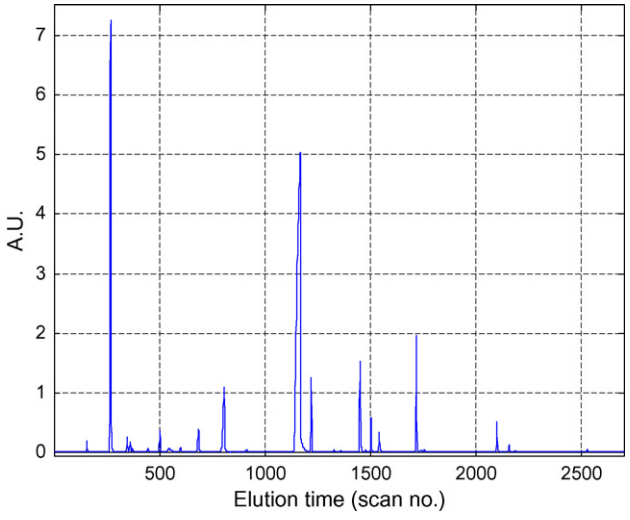


Fig. 3 – Data block #2: GC Elution profiles. Typical chromatogram showing the total ion count (TIC) of one red wine sample.

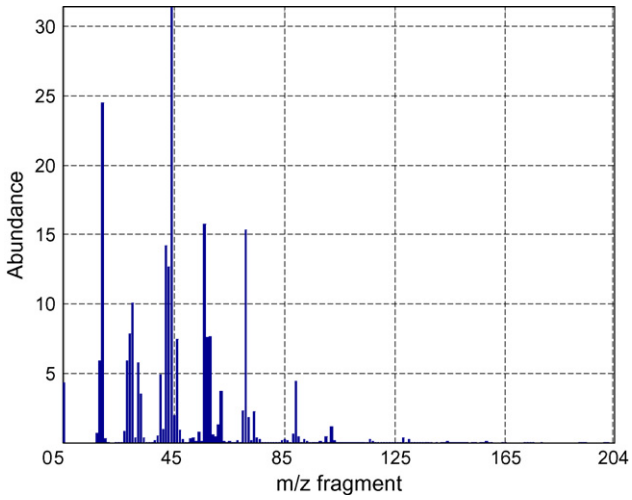


Fig. 4 – Data block #3: GC mass profiles. Typical MS profile of one red wine sample.

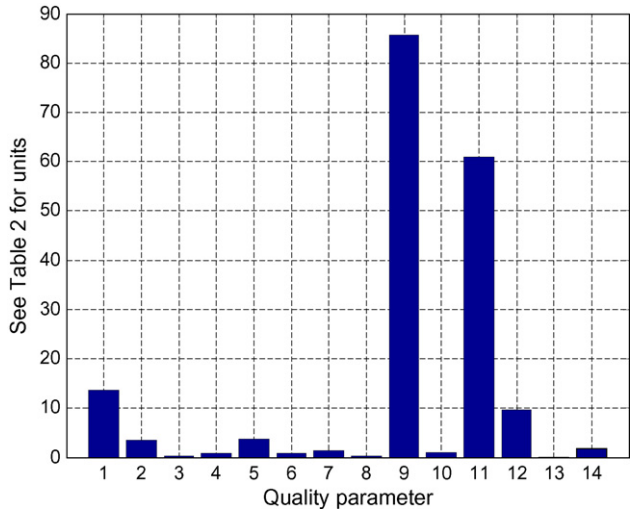


Fig. 5 – Data block #4: IR Quality parameters. Predicted concentration of quality parameters (Table 2) from the FT-IR spectrum of one red wine sample shown in Fig. 6.

$\mathbf{Q}_k$ is the predicted matrix using $\mathbf{Z}_k$.
$\mathbf{Y}_{U,unique}$ is the predicted matrix/vector $\mathbf{Y}_U$ excluding common variation with other $\mathbf{Q}_k$.

Measures describing the variance partitioning:

V^Y variance of values of $\mathbf{Y}$ set to 100%,
V^E unexplained part (residuals, E) of $\mathbf{Y}$ using F PLS factors (% of V^Y),
V^M part of $\mathbf{Y}$ explained from $\mathbf{X}$ using F PLS factors (% of V^Y),
V^C common explained part of $\mathbf{Y}$ using F PLS factors between all predictor matrices—$\mathbf{X}$ and $\mathbf{Z}$ (% of V^Y),
V^U unique part of $\mathbf{Y}$ given as $V^U = 100 - V^C - V^E$ and from Eq. (2).

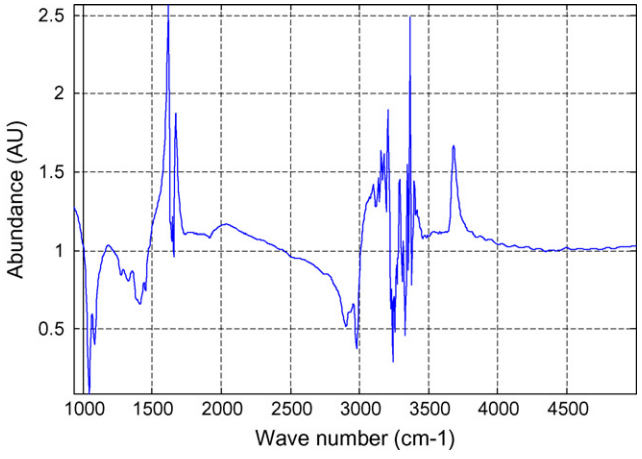


Fig. 6 – Data block #5: IR spectra. Typical FT-IR spectrum of one red wine sample. The water band regions around 1545–1710 cm^{-1} and 2968–3620 cm^{-1} are excluded from the data analysis.

All measures (V) are from calibrated predictions of $\mathbf{Y}$ (number of PLS components estimated from full cross-validation) and the variance of the elements of $\mathbf{Y}$ means the pooled variance of the column variances. In the following variation and variance will be identical terms for the latter.

2.5.2. *Unique variation*

For a particular $\mathbf{X}$-block we seek how much of $\mathbf{Y}$ can be predicted by the given $\mathbf{X}$-block *and*, in addition, how much of this variance is unique to the given $\mathbf{X}$-block taking the $\mathbf{Z}$-blocks into consideration. To find the unique part of the original data the following steps are carried out:

(1) The unique part of $\mathbf{Y}_U$ is found by removing from it variation common with $\mathbf{Q}_k$ variable-wise. Having for example, $k = 1, 2, 3$ a matrix of $[\mathbf{q}_1\ \mathbf{q}_2\ \mathbf{q}_3]$ (denoted $\mathbf{D}$) is obtained containing the predictions of $\mathbf{Y}$ from the additional $\mathbf{Z}_k$ block 1, block 2 and block 3 (the first column of $\mathbf{Q}_k$), respectively. In this case the matrix is size $I \times 3$. This is done for each column in $\mathbf{Q}_k$ in turn.
(2) The first column of $\mathbf{Y}_U$, $\mathbf{y}_U$ is then orthogonalized with $\mathbf{D}$ according to:

$$\mathbf{y}_{U,unique} = (\mathbf{I} - \mathbf{D}\mathbf{D}^{+})\mathbf{y}_U \quad (1)$$

where + indicates the pseudoinverse operator and $\mathbf{I}$ the identity matrix of dimensions $I \times I$.
(3) This results in a vector $\mathbf{y}_{U,unique}$ with only unique variation found in $\mathbf{Y}_U$ for column j.

The unique part of $\mathbf{Y}_U$ is then calculated over all variables/columns to give the unique part, $\mathbf{Y}_{U,unique}$.

As the unique part is found from the explained part of $\mathbf{Y}$, $\mathbf{Y}_U$, and not from $\mathbf{Y}$ itself, the unique part with respect to $\mathbf{Y}$ can be found from:

$$V^U = \frac{var(\mathbf{Y}_{U,unique})}{V^Y} \times 100\% \quad (2)$$

where var means the pooled variance of the column variances. Having found the unexplained part of $\mathbf{Y}$ (V^E) and the unique part of $\mathbf{Y}$ (V^U) from the above equations the common part between all k matrices of $\mathbf{Y}$ can be found from:

$$V^C = V^Y - V^E - V^U \quad (3)$$

and thus $V^M = V^C + V^U$ and $V^U + V^C + V^E = 100$ (see notation above). These measures will be further demonstrated in the next section.

Fig. 7 illustrates the principles of MVP sequentially in a more simplified manner.

2.5.3. *Example of MVP*

An example of how the illustration of these three parts of variation/variance (unexplained, unique and common) can be interpreted is presented using three general data blocks (A–C):

With data block A as $\mathbf{X}$, B as $\mathbf{Z}_1$ and C as $\mathbf{Y}$ we calculate the unique part of data block A explaining C with respect to B.

Then, if A explains 80% of C ($V^M = 80\%$) and the unique part of these 80% with respect to B is 50% we get:

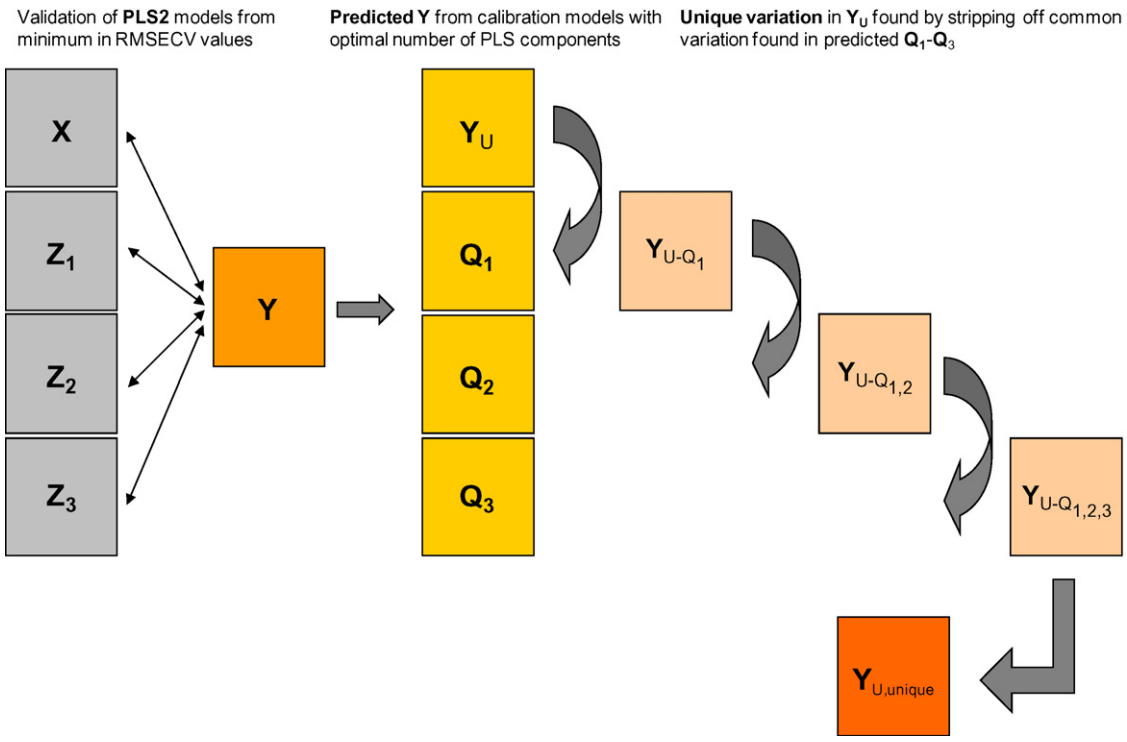


Fig. 7 – Diagram illustrating the multiblock variance partitioning approach. The sequential removal of common variation is shown to make the visualization more clear. The correct mathematical equations are given in Section 2.5, where further details of MVP can also be found. $Y_{U,unique}$ is the same as $Y_{U-Q1,2,3}$.

Unexplained variation (V^E) = 20%.
Common variation (V^C) = 40%.
Unique variation (V^U) = 40%.

Consequently, unexplained, common and unique parts (variations) sum up to 100%. This relationship and the partition of the total information can be easily visualized in a bubble plot (Fig. 8). The complete bubble represents the 100% of information in **Y**, while each kind of variation (unexplained, common and unique) is represented with a different color, making the interpretation of the relationship between the data blocks easier.

With Fig. 8 as an example we can define what unexplained, common and unique variation are in words:

Unexplained variation is the percentage of information of **Y** that **X** cannot explain (i.e. **X** is not related to this part of the information in **Y**); the larger this unexplained variation is, the less the two data blocks are related. The unexplained variation will thus consist of noise as well as systematic information in **Y** that is not related to **X**.

Common variation is the percentage of information of **X** that is also found in any of the **Z**-blocks $\mathbf{Z}_1$,..$\mathbf{Z}_K$ when predicting the same **Y**-block.

Unique variation is the percentage of information in **X** that is not related to variation found in the **Z**-blocks $\mathbf{Z}_1$,..$\mathbf{Z}_K$ when predicting the same **Y**-block.

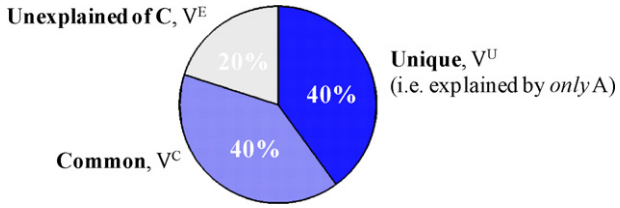


Fig. 8 – Illustration of unexplained, common and unique variation found using the three data blocks A–C mentioned in the text.

2.6. *Software*

Chemometric models were calculated in MATLAB® version 7.3 (R2006b) (The MathWorks Inc., Natick, MA, USA) by means of the PLS toolbox (Eigenvector Research, Inc., Manson, Washington). In-house functions (www.models.life.ku.dk) were used for the MVP approach to calculate the unique and common variation.

3. Results and discussion

3.1. *How to use and interpret MVP*

The five considered data blocks (Table 3) have all been used iteratively as the **Y** data block and among the remaining four; one as the specific **X** data block and the other three as **Z** data blocks according to the principles of MVP explained before. In Fig. 9 the unexplained/explained part, the common part and the unique part are shown as percentages of the total variation in the original **Y** data block.

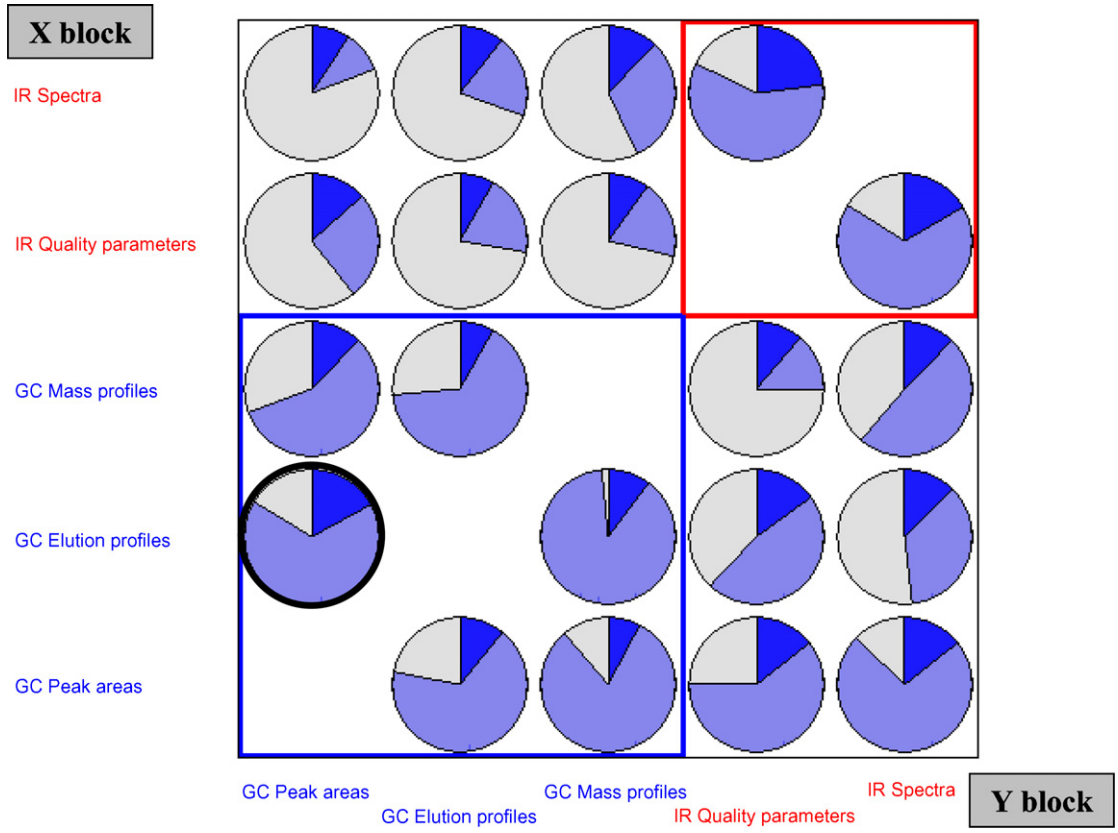


Fig. 9 – Unexplained part (light grey), common part (light blue) and unique part (dark blue) in percentages of the variation in Y—see Fig. 8 for depiction of these parts. The blue letters and lines refer to the GC related data blocks and the red to the IR related data blocks. Example (fat circle ◯): GC peak areas as Y (response) and GC elution profiles as predictor X—unexplained variation, 16.3%; common variation, 63.3% and unique variation, 20.4%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

In Fig. 9 the main results comparing variation within and between analytical techniques are shown. To make clear how the figure could be interpreted some intuitive guidelines and findings are presented:

A single row describes how one data block explains the others. The first row with IR spectra as **X**-block visualizes how this data block is able to predict the others. For example, IR spectra are more related to the other IR technique than the GC representations (larger explained variance part).

A column visualizes how this data block is predicted, i.e. how well the other data blocks are able to predict it. For example – in the first column – the two other GC data representations are able to describe GC peak areas better, as expected (larger explained variance part). If a column is compared with its associated row (for example the fifth row and first column, both related to the GC peak areas), the figure shows that the GC peak areas are better at *describing* (predict) than at being described (predicted)—since the explained part (common plus unique part) are larger on the fifth row than on the first column).

A larger unique part compared to the common part indicates that the predictive data block (**X**) and the data block being predicted (**Y**) have a lot in common (if the unique part is large) and that a large part of this communality is unique to the predictive data (**X**)—e.g. GC mass profiles as **X** and IR quality parameters as **Y**. On the other hand, a small unique part and a large common part—e.g. GC mass profiles as **X** and GC elution profiles as **Y**-means that these data blocks have a lot in common, but that much of the explained variance is also found in one or more of the other data blocks. If both parts are small then the unique and common parts are more difficult to interpret, but overall it can be concluded that the two data representations in this case only contain very little related information.

The explained parts (unique plus common part) are in general larger when evaluating two data representations from the same analytical technique (GC or IR). This indicates that data representations are more similar when originating from the same analytical technique.

When considering the relationship between GC and IR representations, the trend is that the unique and common parts decrease due to less commonality between data blocks. Some deviating trends are however found here: if GC peak areas (or GC elution profiles) are predicting the IR data representations a fair amount of the IR data is explained (up to 87% for GC peak areas predicting IR spectra) and thus, the GC peak areas and the IR spectra seems to have more in common than when

these data are used to predict or being predicted by other data blocks.

The two most similar data representations (highest explained variation part) are found when GC elution profiles predict the GC mass profiles. Predicting GC mass profiles from GC elution profiles provides that 99% of the variation in GC mass profiles is explained. This confirms that GC data representations are rather similar, and also that the GC elution profiles might be used instead of the normal fingerprint (GC mass profiles) in, e.g. classification studies. In the GC mass profiles there is information from the fragments of the eluting compounds, but using electron ionization (EI), the fragmental fingerprint is very often too complex to enable any identification of important analytes. Changing the ionization technique to a softer method (i.e. CI–providing less fragment and higher *m*/*z* values with significant intensities) more structural information would be preserved.

The comparison of all five data representations shows that no single data block contains variation not explained by (common with) one or more of the other data representations when predicting the same **Y** data—if this was the case then the explained variation part should be equal to the unique part in Fig. 9. This is not surprising as one or more related data representations (either GC or IR) are included in all circles shown in Fig. 9.

Fig. 9 suggests that GC–MS and FT-IR data are complementary analytical techniques as they each bring information that cannot be found from the other technique. The MVP approach shows that some pieces of information are still redundant and that some variables from individual techniques could hold the same information (i.e. be correlated). The MVP method presented here does not provide an option to evaluate the influence of individual variables in individual PLS models. To exclude irrelevant or noisy variables preliminary variable selection must be conducted on each of the initial PLS models. However, as the PLS models have been optimized in cross-validation with respect to number of components to include only the relevant variables should be influencing the established relationship between the **X**- and **Y**-block.

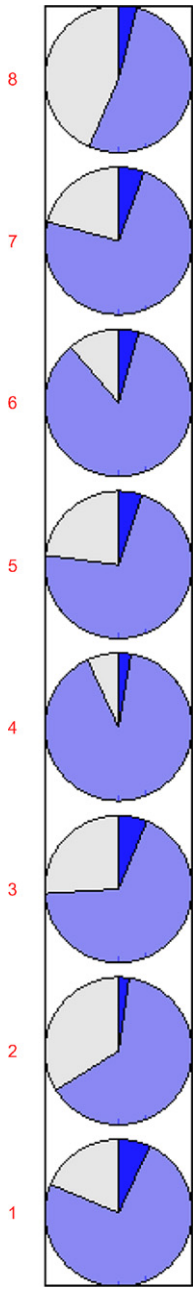


Fig. 10 – MVP for sensory data evaluating the assessor effect. Each predictor block is the assessor scores of six pork patties in duplicates evaluated the meat flavor in a time-intensity study [22]. The Y-block is the average over all assessors, which is the normal procedure for data pre-treatment sensory science.

3.2. *Other ways to use MVP*

MVP is a generic tool and can be useful in many typical data analysis problems. To exemplify this, a few selected examples are given in the following to provide inspiration for other ways to use MVP.

3.2.1. *Process control*

The MVP technique can be useful when monitoring a process at different times and at different positions from raw material to final product. Fixing the **Y**-block to the final product, data from each time and position can be evaluated with respect to unique information taking all other times and positions into consideration. In this way it can be evaluated if a certain measurement is redundant (small unique part) or if special focus is needed (large unique part).

3.2.2. *Sensory data—assessor difference*

For sensory data the assessors are often regarded as replicates in evaluating the same product. Differences between assessors can occur, e.g. due to different use of the scale for individual attributes (i.e. similar profile over all attributes but in different magnitudes). This 'offset' can be corrected by a simple mean centering approach [21]. However, if the assessors have different profiles for the same sample then different sample descriptions will be included in the model. To avoid

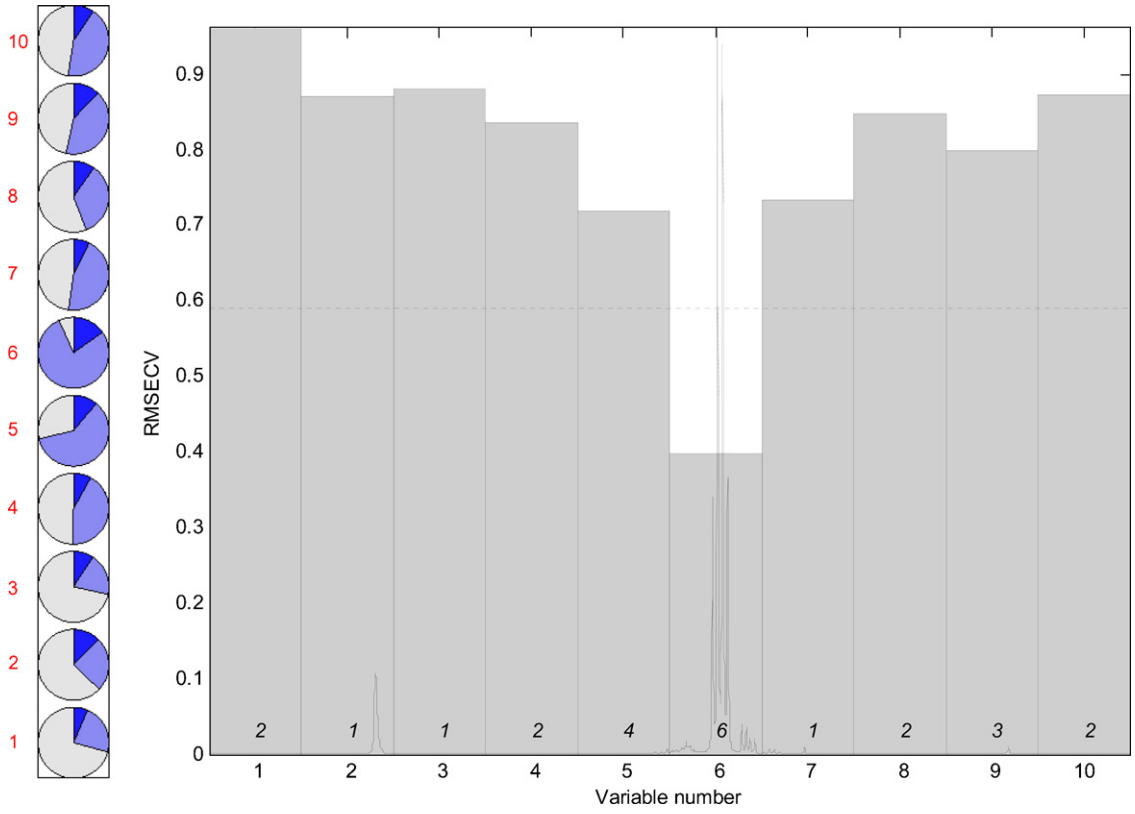


Fig. 11 – Right: iPLS predicting the ethanol content of red wines samples from part of original NMR spectra [24] each divided into 10 intervals. Left: MVP using the same intervals as X- and Z-blocks in turns and ethanol as Y-block. The optimal number of PLS factors is found by the minimum value of RMSECV for the 10 local PLS models using 10 PLS factors in model. Dotted line represents the RMSECV value from a model with all intervals.

this difference between assessors the average over the assessor scores is normally used in subsequent data analysis. The use of advanced multiway models to correct for this difference has been put forward in [21].

With different assessor profiles MVP could be a useful method to see which assessors bring common information (i.e. are good replicates) and which assessors are different/outliers (i.e. low explained variance of **Y**) when predicting the same **Y**-block. The **Y**-block should then be the averaged assessor scores.

In Fig. 10 an example of eight assessors predicting the same **Y**-block is shown. The data is kindly provided from a time-intensity study of pork patties evaluating the meat flavor over time [22].

Fig. 10 shows that the assessors used for the meat flavor study performs quite different when compared to the average assessor performance (**Y**-block). E.g. the assessor four performance is very similar to the average performance, whereas assessor eight shows a deviation compared to the average. The MVP results reveal that assessor eight has a different profile (i.e. different time-intensity profile) compared to the average profile. The sensory data presented here is an extreme case as time-intensity profiles are known to differ to some extent between assessors. But the principles of applying and interpreting the MVP results are the same and can be used for sensory date with normal sensorial attributes. In [21,22] the use of advanced multiway methods to handle these differences are presented.

3.2.3. *With iPLS—as a complementary technique*

MVP can also be used as a complementary technique with interval PLS (iPLS) [23]. iPLS is originally designed to evaluate different regions of spectra to find the region giving best prediction (smallest prediction error) of a certain response. This can be done from individual regions of user-defined sizes or combining the best regions in a forward or backward selection step. iPLS focuses on prediction of the response variable(s) and thus, it will not tell whether the information found in the best predictive region can be found elsewhere in the spectra. Combining iPLS with MVP can enhance the understanding of the prediction and guide the combination of specific regions.

In the following an example predicting the alcohol concentration in red wines from NMR data kindly provided by [24] will be shown using both iPLS and MVP (Fig. 11).

The best predictive interval is the data block holding the major peaks (interval 6), which is also providing the largest calibrated explained **Y** variance. From iPLS it can be concluded that a prediction of the ethanol content in wines would be best using interval 6. However, MVP complements iPLS to give an idea of the uniqueness of the variance found in this interval. It can be seen than approximately 20% the explained variance of **Y** is due to variance only found in this interval. Thus some

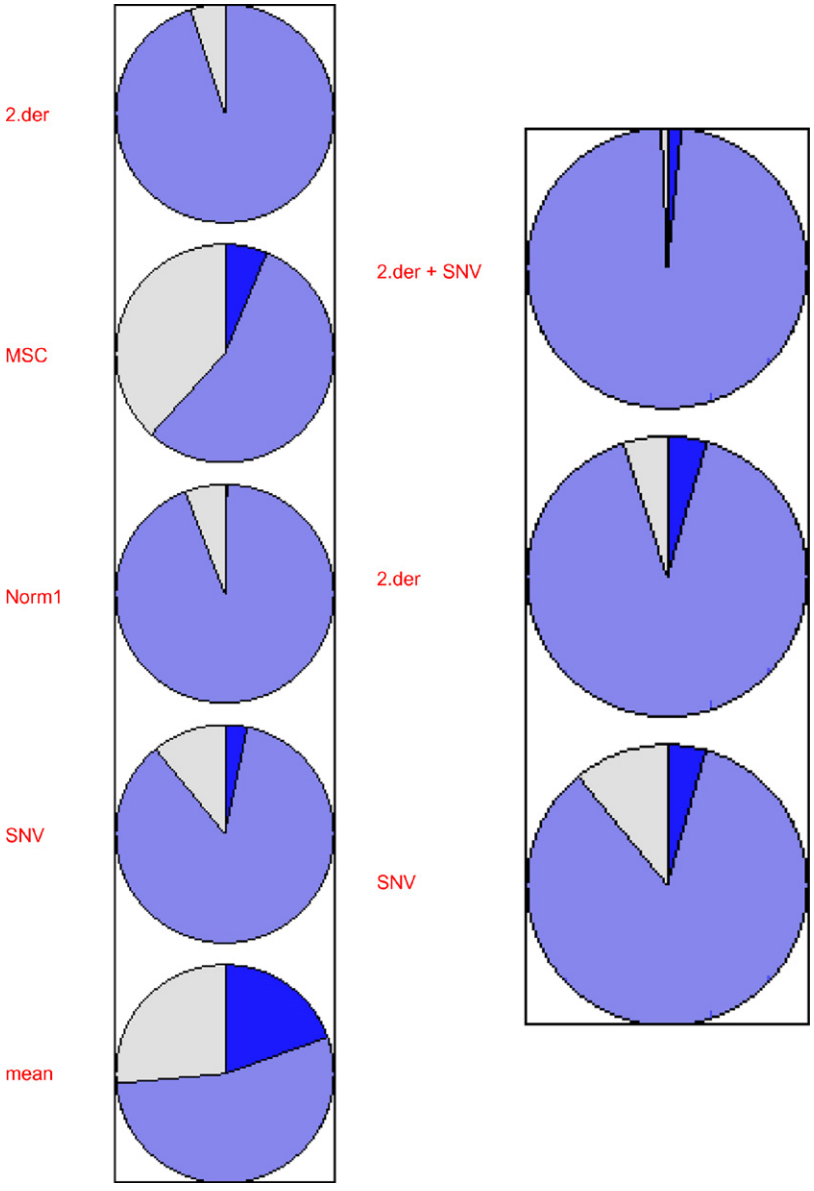


Fig. 12 – MVP for different preprocessing method. Each predictor block is the preprocessed data using either mean centering alone, SNV, Norm1, MSC or second derivative. The Y-block is the gluten content [25]. All blocks have been mean centered after preprocessing with the indicated method and one PLS component has been used in all models as suggested in [25].

common information is located in the surrounding intervals as well. This advocate for a combination of intervals as the best predictive block, but with the interval sizes given here no combination gave better prediction of the alcohol content. Smaller intervals and a more detailed investigation of combining proper intervals could be an alternative to get a better prediction, but this was not pursued here.

3.2.4. *Preprocessing steps*

Another area where MVP could be used is when several data blocks must be compared and where it makes sense to evaluate them at the same time instead of in sequential steps. E.g. comparing different preprocessing techniques in predicting a certain **Y**-block one can estimate both the influence of each individual technique and also the commonality between techniques. This approach is often tested using a certain prediction error (i.e. RMSECV/RMSEP), but here quantitative measures of common and unique parts can be achieved. An example is given from the data presented in [25], where several preprocessing methods were applied for near-infrared transmission data. Data consist of mixtures of two powder types; wheat gluten and wheat starch with the response being the gluten fraction. An MVP analysis of these data is presented in Fig. 12.

Here we have applied different preprocessing methods and from Fig. 12 it can be seen that the normalization ($V^M = 94\%$) and second derivative ($V^M = 95\%$) perform very well and they seem to contain the same phenomena in the block after preprocessing (no unique part). This was confirmed with MVP on only these two data blocks—results not shown. If mean centering is performed as the only preprocessing method, a large

unique part is present compared to the other methods. This indicates that the mean centered block cannot describe the **Y**-block structure as adequately as the blocks treated with second derivative and normalization. No preprocessing method can make the spectral data fit the **Y**-block perfectly, but combining two or more blocks might improve this. The left part of Fig. 12 suggests that the second derivative and normalization are good choices, but as these blocks describe the same phenomena this would not improve the model (results not shown). However, SNV is seen to contain additional information not found in the two best blocks and may thus lead to improved predictions if combined with one of the best blocks. This is exemplified in Fig. 12 (right part). The changed unique and common parts (same explained variation of **Y**) are due to normalization and mean centering being left out of this MVP calculation. Now, also the second derivative treated block holds a unique part and thus, extra information about the **Y**-block will be included when combining the two preprocessing methods resulting in an increased explained variation of the **Y**-block.

4. Conclusion

A new method for evaluating the variation found in different data blocks has been presented. The method called multiblock variance partitioning works by establishing local PLS models between predictor blocks and a common response block to be able to find unique and common variation in the predictor blocks. MVP works on individual predictor blocks and thus, no scaling issues needs to be considered, which is often problematic in multiblock analysis. The MVP approach brings two new quantitative measures; a unique and a common part. The unique part is the part of variation only found in one specific predictor block, whereas the common part is the variation found in one or more of the other blocks.

It has been shown that MVP can be used to evaluate the redundancy in multiple blocks. In sensory analysis a large redundancy between assessor blocks (large common part) was shown to be an indication that the blocks can be used as replicates. MVP was also advantageous when single preprocessing methods showed good potential, but a combination of the two proved to be even better due to unique parts found in the individually treated data blocks.

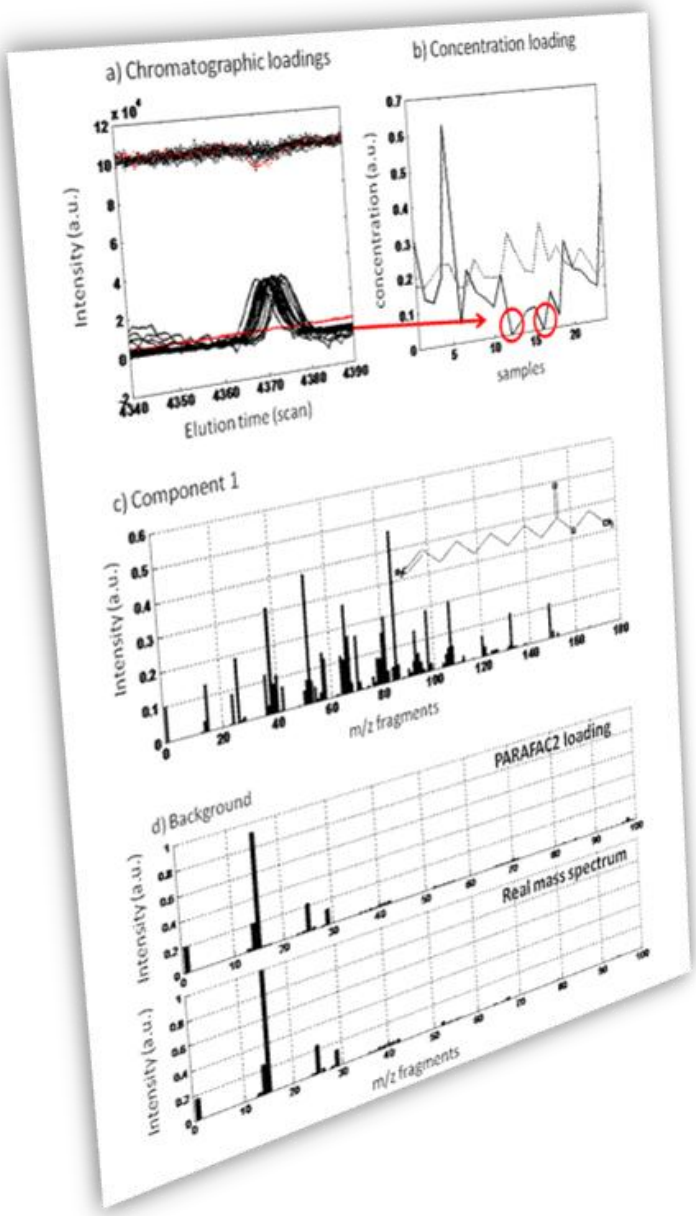
As a multiblock technique, MVP handles individual predictor blocks separately in optimized PLS models and thus, no scaling-of-blocks issues need to be considered. The MVP approach has been shown to be an easy and powerful method that can be applied as an additional tool in very diverse multiblock data analyses.

REFERENCES

[1] S. Wold, K. Esbensen, P. Geladi, Chemometr. Intell. Lab. Syst. 2 (1987) 37.
[2] S. Wold, M. Sjostrom, L. Eriksson, Chemometr. Intell. Lab. Syst. 58 (2001) 109.
[3] A. Berglund, S. Wold, J. Chemometr. 13 (1999) 461.
[4] C.C. Felicio, L.P. Bras, J.A. Lopes, L. Cabrita, J.C. Menezes, Chemometr. Intell. Lab. Syst. 78 (2005) 74.
[5] M. Tenenhaus, V.E. Vinzi, J. Chemometr. 19 (2005) 145.
[6] A. Hoskuldsson, K. Svinning, J. Chemometr. 20 (2006) 376.
[7] J.A. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemometr. 12 (1998) 301.
[8] J.A. Westerhuis, A.K. Smilde, J. Chemometr. 15 (2001) 485.
[9] S. Wold, N. Kettaneh, K. Tjessem, J. Chemometr. 10 (1996) 463.
[10] J. Gabrielsson, H. Jonsson, C. Airiau, B. Schmidt, R. Escott, J. Trygg, J. Chemometr. 20 (2006) 362.
[11] I. Måge, Modelling and optimisation of industrial processes with raw material variation, PhD Thesis, Norwegian University of Life Sciences, September, 2006.
[12] C.J. Bevin, A.J. Fergusson, W.B. Perry, L.J. Janik, D. Cozzolino, J. Agric. Food Chem. 54 (2006) 9713.
[13] X. Capron, J. Smeyers-Verbeke, D.L. Massart, Food Chem. 101 (2007) 1585.
[14] O. Gurbuz, J.M. Rouseff, R.L. Rouseff, J. Agric. Food Chem. 54 (2006) 3990.
[15] S.M. Rocha, P. Coutinho, A. Barros, I. Delgadillo, M.A. Coimbra, J. Chromatogr. A 1114 (2006) 188.
[16] I. Arozarena, A. Casp, R. Marín, M. Navarro, J. Sci. Food Agric. 80 (2000) 1909.
[17] Y. Kotseridis, A. Razungles, A. Bertrand, R. Baumes, J. Agric. Food Chem. 48 (2000) 5383.
[18] J. Forshed, H. Idborg, S.P. Jacobsson, Chemometr. Intell. Lab. Syst. 85 (2007) 102.
[19] T. Skov, R. Bro, Anal. Bioanal. Chem. 390 (2008) 281.
[20] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemometr. 20 (2006) 484.
[21] R. Bro, E.M. Qannari, H.A.L. Kiers, T. Næs, M.B. Frøst, J. Chemometr. 22 (2008) 36.
[22] H.C. Reinbach, L. Meinert, D. Ballabio, M.D. Aaslyng, W.L.P. Bredie, K. Olsen, P. Møller, Food Qual. Prefer. 18 (2008) 909.
[23] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413.
[24] F.H. Larsen, F. van den Berg, S.B. Engelsen, J. Chemometr. 20 (2006) 198.
[25] H. Martens, J.P. Nielsen, S.B. Engelsen, Anal. Chem. 75 (2003) 394.

PAPER IV

Amigo, J.M., Skov, T., Coello, J., Maspoch, J., and Bro, R. **2008**. Solving GC-MS problems with PARAFAC2. *Accepted for publication in Trends in Analytical Chemistry - TrAC.*



SOLVING GC-MS PROBLEMS WITH PARAFAC2

José Manuel Amigo*[a], Thomas Skov[a], Jordi Coello [b], Santiago Maspoch [b], Rasmus Bro[a]

[a] Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

[b] Departamento de Química, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

*Corresponding author: jmar@life.ku.dk

Accepted for publication in Trends in Analytical Chemistry – TrAC

Abstract

Gas Chromatography-Mass Spectrometry (GC-MS) is an important technique for identification and quantification of analytes in multifactor systems. Nevertheless, the experimental sources of variability related to GC-MS (e.g. column and flow meter ageing, changes in certain characteristics or properties of the stationary phase, and changes in temperature, experimental conditions or preparation of the standards, chemicals, etc) may cause variations in elution time, baseline drifts, unexpected overlapping of peaks, non-gaussian peaks, etc. Several approaches have been proposed to handle these problems, with the standardization of peak areas using internal standards being one of the most efficient techniques. However, such a solution is not sufficiently versatile when deviations from ideality are more pronounced.

Since a mass spectrum can be obtained at each elution time during a chromatographic separation, GC-MS data of several samples can be considered a three-way structure. PARAllel FACtor analysis 2 (PARAFAC2) is a model capable of handling three-way data and, unlike the PARAFAC model, does not assume that the elution profiles of each factor are invariant across samples. This coupled with the uniqueness properties of PARAFAC2 allows PARAFAC2 to solve several problems derived from experimental conditions in GC-MS datasets.

This paper is aimed at showing the potential of PARAFAC2 for solving common GC-MS problems. GC-MS data of wine samples are used to illustrate the solutions.

Keywords: PARAFAC2; hyphenated Chromatography; GC-MS; elution time shifts; baseline drifts; overlapping peaks

Nomenclature

A, B, C, D, E, X	Two-way data and parameter matrices
COW	Correlation Optimized Warping
F, I, J, K	dimensions of matrices and arrays
f, i, j, k	indexes of the dimensions of the matrices
FID	Flame Ionization Detector
GC-MS	Gas Chromatography-Mass Spectrometry
HPLC-UV	High Performance Liquid Chromatography-Ultra Violet detector
LC-MS	Liquid Chromatography-Mass Spectrometry
MCR-ALS	Multivariate Curve Resolution-Alternating Least Squares
PARAFAC	PARAllel FACtor analysis
PARAFAC2	PARAllel FACtor analysis 2
SSE, SSX	sum of squares of residuals and data, respectively
$\mathbf{X}^T$	transpose of matrix X
$\underline{\mathbf{X}}$	three-way array
x_i, $\hat{x}_i$	intensity for m/z equal to *i* for the experimental spectrum and to the calculated loading respectively

1. Introduction

1.1. Common problems in GC-MS datasets

Gas Chromatography (GC) is a very powerful separation technique for multifactor mixtures of analytes. An ideal chromatographic experiment leads to perfect separation of these chemical analytes. The resolving power can be increased when a Mass Spectrometer (MS) is coupled to the chromatograph, since the mass spectrometry dimension provides additional resolving power compared to traditional detectors such as the Flame Ionization Detector (FID)[1]. The high sensitivity, the low limit of detection, the possibility of analysing a great amount of analytes and identifying these using the mass spectra, makes GC-MS one of the most widespread analytical techniques in many scientific fields. Several excellent reviews have recently been published showing the possibilities and limitations of GC-MS in biological matrices [2], in geological and forensic research [3, 4], environmental matrices [5, 6], proteomics [7-10] and metabonomics [11].

Despite the above, analysis of GC-MS data is sometimes hampered by different problems mainly derived from the chromatographic separation and/or mass spectral measurements. Sometimes it is not possible to achieve perfect separation, either because of the complexity of the samples or because faster chromatographic runs are preferred. Also, problems with drift in the baseline, changes of shapes of the peaks and shifts in the elution times may decrease the quality of the final result of the analysis [12]. Traditionally, the use of, for example, one or more standards help to control and locate some of these problems. However, the capability of the internal standards is limited in more severe cases of baseline drifts and change of shapes of the peaks.

Due to the complexity of the samples often measured using GC-MS, standard data analytical tools from instrument vendors are insufficient for extracting all the relevant information from such data. A more versatile methodology is needed for the analysis of the datasets providing results that can be plotted in an easy and comprehensive way.

Several mathematical methods have already been proposed to correct for undesirable phenomena introduced during the chromatographic run [11, 13-18]. Unfortunately, the use of advanced data analysis focusing on detailed analysis of GC-MS data is not widespread. For example, the Correlation Optimized Warping (COW) algorithm has shown promise in handling elution time shifting due to its peak shape and area preserving properties [19]. Skov et al. [20] put

forward an interesting approach for how to choose the best conditions for applying the COW algorithm in correcting elution time shifts while preserving peak shapes However, COW does not allow correcting artefacts such as varying baseline or severe overlapping peaks and can introduce artefacts if peak shape changes are observed across samples.

1.2. Structure of GC-MS data

When a GC-MS measurement is performed, a two-way matrix is obtained for each sample (Figure 1a). The matrix corresponding to the k'th sample $\mathbf{X}_k$ is of dimensions ($I \times J$) where the columns represent each of the length I mass spectra for each of the J elution times (rows). A three-way structure is obtained when measurements from a set of samples are obtained and stacked (Figure 1b). The cube of data will be $\underline{\mathbf{X}}$ ($I \times J \times K$).

1.3. Three-way models

During the last ten years, there has been an increased focus on taking advantage of this three-way structure of the GC-MS data, or of any other hyphenated chromatographic technique (like, LC-MS or HPLC-UV). This has been made possible by the improvement in data acquisition and the power of personal computers. Models such as, Multivariate Curve Resolution-Alternating Least Squares in the unfolded way (MCR-ALS) [21], Generalized Rank Annihilation Method (GRAM) [22] and PARAllel FACtor analysis (PARAFAC) [23] have been used.

The first general solution for handling the overlapping peaks problem was the Generalized Rank Annihilation Method (GRAM), by Sánchez and co-workers [24, 25]. This method was a generalization of the method Rank Annihilation Factor Analysis (RAFA) of Ho and co-workers [26], mainly used for quantification purposes. The main feature of GRAM is that it works even with only two samples. The first sample usually holds the chromatographic profile of a mixture containing the known standard(s) and the second one contains the analytes whose profiles are overlapped.

RAFA and GRAM have a serious drawback: they can only be used in cases of a single standard and a single unknown mixture. Hence, the results rely heavily on the quality of the two samples and no additional benefits can be gained from having more samples. To overcome this problem, Sánchez and co-workers developed the Direct Trilinear Decomposition algorithm (DTLD)

[27]. This algorithm is an extension of GRAM but for more than two samples. This algorithm decomposes the data into three matrices, containing the elution profiles, the mass spectra and quantitative information about each of the components in each sample. Both GRAM and DTLD are efficient when the data are well approximated by the underlying model, but they are often too sensitive to minor deviations e.g. caused by retention time shifts.

The parallel Factor Analysis (PARAFAC) model [28, 29] is probably the most widespread multi-way method in chromatography. PARAFAC decomposes the three-way array into three two-way matrices, called loading matrices, one for each mode. For the sample mode, the loading matrix is usually referred to as a score matrix and holds the relative concentration of each chemical compound in the sample if the model successfully separates the individual chemical compounds into individual PARAFAC factors. The elution mode loading matrix correspondingly holds the estimated elution profiles of each analyte and the last mode, the estimated mass spectra of each factor.

The major problem with GRAM, DTLD and PARAFAC is related to the intrinsic structure of the data array that these models can handle. The models all assume that the elution profiles and the mass spectra remain invariant (i.e. same profile) for all the samples, i.e. the data set is low-rank trilinear and that data from each analyte has a rank-one contribution to the data [30, 31]. In an ideal situation, this assumption will be valid. Nevertheless, this assumption is often far from the reality due to different sources of variability in the elution time mode of GC-MS data. Consequently, trilinearity of the data is disturbed especially due to the changes in the elution time mode [32] (Table 1).

One method that allows a certain freedom in the chromatographic profiles is the MCR-ALS [33, 34]. Here the analysis is performed unfolding the three-way array in the chromatographic direction keeping the mass spectral dimension intact. The main advantage of MCR-ALS is that it can also handle non-trilinear data. Nevertheless, MCR-ALS is highly dependent on the initial estimates of input parameters (elution and mass spectral profiles) and the three-way information is lost in the unfolding stage. This implies that it is not possible to recover estimates of the signal from the underlying analytes unless certain requirements are fulfilled.

1.4. PARAFAC2 as an alternative to PARAFAC

Like PARAFAC, the PARAFAC2 [35, 36] model also decomposes three-way data arrays into loading matrices, but the main difference is that PARAFAC2 does not impose as strong restrictions on the data structure [31]. PARAFAC2 does not assume that the shape (or even length) of the elution profile of an analyte is the same in each sample (Table 1). This property has allowed the use of the PARAFAC2 model in such diverse fields as sensor based data [37], semiconductor production [38], HPLC-DAD [39], kinetic spectrophotometric analysis [40-43], modelling of flavour release [44] and analytical dilution systems [45, 46]. Even though PARAFAC2 allows the elution profiles to differ in shape in different samples, it still possesses uniqueness properties that are very similar to PARAFAC [47]. This means that if a successful model is obtained, PARAFAC2 can separate mixture data into the contributions (concentrations, elution profiles and mass spectra) of the underlying analytes directly. PARAFAC2 could be an ideal technique for modelling GC-MS data as the model allows obtaining one elution time profile for each factor of the sample taking into consideration that each analyte has a mass spectrum that is consistent across all samples. Excellent examples of the use of PARAFAC2 in GC-MS can be found. For example, in the work of Cruz et al. calibration models were developed with PARAFAC2 for the quantification of five hormonal growth promoters [32] and of several steroidal hormones [23], and similar results obtained compared to an internal standard calibration in terms of the figures of merits. Also in the work of Hibbert et al. [48] PARAFAC2 was applied to classify different weathered oils. Nevertheless, as Skov et al. [49] stated, further investigations are needed to fully establish the usefulness of PARAFAC2 in modelling GC-MS data.

This paper will present a comprehensive overview of several problems in GC-MS such as severe shifts in retention time, overlapping peaks, baseline drifts and low intensity peaks and how the use of PARAFAC2 can handle these. The paper will also highlight the possible limitations using PARAFAC2.

2. PARAFAC2 as a model of GC-MS data

Three-way chromatographic data can be arranged in such a way that the first mode refers to the mass channels, the second mode to the elution times and the third mode to the samples (Figure 1b).

The PARAFAC2 model allows that every sample can have its own distinct set of elution time loadings [50]. Each sample (each slab of $\underline{\mathbf{X}}$) $\mathbf{X}_k$ ($I \times J$) is modelled as indicated in equation 1 [50] (Figure 2).

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k(\mathbf{B}_k)^{\mathrm{T}} + \mathbf{E}_k \quad ; k = 1,..,K \tag{1}$$

where $\mathbf{X}_k$ ($I \times J$) represents the chromatographic run related to the kth sample. $\mathbf{A}$ ($I \times F$) holds the resolved mass spectra of the F analytes. $\mathbf{D}_k$ ($F \times F$) is a diagonal matrix that holds the k'th row of the sample mode loading matrix $\mathbf{C}$. Each element in this row is the relative concentration of analyte f. The matrix $\mathbf{B}_k$ ($I \times F$) ideally holds an estimate of the individual elution profiles for each of the F factors. The matrix $\mathbf{E}_k$ represents the residuals.

An important constraint in the PARAFAC2 algorithm is that the cross-product of $\mathbf{B}_k$ has to be constant over all the samples (equation 2). This constraint is needed for obtaining uniqueness. The implication of this cross-product constraint is that the elution profiles in different experiments may differ e.g. due to shifting, but should still be somewhat similar.

$$(\mathbf{B}_1)^{\mathrm{T}}(\mathbf{B}_1) = (\mathbf{B}_2)^{\mathrm{T}}(\mathbf{B}_2) = \ldots = (\mathbf{B}_k)^{\mathrm{T}}(\mathbf{B}_k) \tag{2}$$

A visualization of the decomposition of the three-way array $\underline{\mathbf{X}}$ can be seen in Figure 2. Detailed information about the model and algorithms can be found in [50].

3. Experimental

The data used here consist of GC-MS landscapes of 24 samples of red wine of different origin (Figure 3) [51]. A Dynamic headspace gas chromatography mass spectrometry (**HS-GC-MS**) system was used to measure the aroma profile of the wine samples. More information about the collection and the experimental set-up of the samples can be found in reference [51].

Typical GC-MS problems are reflected in the data such as severe changes in the shape of several peaks, severe shift in elution time, low intensity of peaks, baseline drift and overlapping peaks. Five areas of the data set have been selected that reflect these problems (solid squares in Figure 3). These areas have been enlarged in Figure 3.

3.1. Software

Currently software for computing PARAFAC2 models is available from two sources. The first one can be freely downloaded from the web [52]. The second one can be obtained from Eigenvector Research and is implemented in PLS_Toolbox v. 4.1 [53]. Both algorithms work in MatLab (The Mathworks, Inc. 2008). In this work the PARAFAC2 algorithm from PLS_Toolbox v. 4.1 was used.

3.2. Data analysis

From an analytical point of view, the estimated mass spectra as well as their corresponding concentrations cannot be negative, so non-negativity constraints were imposed for sample and mass spectral loadings. In the PARAFAC2 algorithm available and used in this work [53] it is not possible to impose non-negativity constraints on the elution profiles even though this would also be appropriate.

The selection of the correct number of factors is the most important model aspect in the application of PARAFAC2. Because of the nature of the model, all parameter estimates (e.g. estimated elution profiles) depend on each other and therefore, the correct number of factors must be used. Otherwise, the parameters will at worst be chemically meaningless even though they are mathematically uniquely determined. In this paper and in chromatographic analysis in general, it is advisable to resolve chromatographic peaks individually hence work on a small window/region of the elution time dimension. That way, the quality of the models will be better because indirect and direct correlations between different compounds do not affect the modelling and because the problem of choosing the number of factors is typically simplified significantly. In the following, the number of factors was determined by assessing the explained variance for the model (Eq. 3) as well as scrutinizing the appearance of the parameters and the residuals. The explained variance was calculated as follow:

$$\%\,\mathrm{var} = 100 \times \left(1 - \frac{SSE}{SSX}\right) \tag{3}$$

where SSE is the sum of the squares of the residuals and SSX is the sum of the squares of the elements of the three-way array. Furthermore, visual interpretation of the results (such as residuals) was used to guide the selection.

All the obtained mass spectral loadings were scaled in order to compare with experimental mass spectra. The experimental mass spectra were obtained from a run of the pure analyte if this was identified from the chromatographic software. Otherwise an average of all mass spectra from all scans constituting an individual peak was used. The comparison was carried out by calculating the similarity between the experimental mass spectrum and the calculated loading. This was calculated as follow:

$$fit(\%) = 100 \times \left(1 - \sqrt{\frac{\sum (x_i - \hat{x}_i)^2}{\sum x_i^{\,2}}} \right) \tag{4}$$

where x_i corresponds to the intensity for m/z equal to i for the experimental spectrum and $\hat{x}_i$ corresponds to the appropriate loading. The Wiley Library for identification of analytes in ChemStation [54] was used to identify the structure of each analyte according to the experimental mass spectrum and the mass spectral loading obtained with PARAFAC2.

4. Results and discussions

4.1. Large shifts in the elution time

The peaks in region "a" of the chromatogram (Figure 3a) represent the classical problem of shifts in the elution time. This is a well defined peak, with high intensity (high signal-to-noise ratio). The peak of one sample is severely shifted but also all the remaining ones are shifted**.**

Results obtained with PARAFAC2 can be observed in Figure 4. In this situation is easy to see that the PARAFAC2 model has to be developed with only one factor. As expected, PARAFAC2 is able to model the chromatogram of the peak for each sample (Figure 4a). Figure 4b shows the relative concentration loading obtained for each sample. This concentration loading value is an estimate of the area under each chromatographic peak. It can be used for calibration models for the analyte of interest. The mass spectrum obtained with PARAFAC2 perfectly matches with the one obtained experimentally. The similarity is 99.9%. In Figure 4c the comparison between the mass spectrum obtained with PARAFAC2 (upper) and the real mass spectra (bottom) is presented. This peak has been identified as acetic acid, ethyl ester [54].

4.2. Overlapping and shifted peaks

The second region selected appears between scan 2635 and 2695. Looking at the enlarged figure (Figure 3b) it is difficult to assess the number of analytes that are present because of the overlaps, shifts and baseline drift.

At first, there seems to be only two slightly overlapping peaks. The explained variance of a PARAFAC2 model with two factors is 97.25% which is on the low side for data such as these (Table 2). The residuals of this model present trends that may indicate that two factors are not enough to explain the systematic variation in the data probably due to a baseline effect reflected in the background mass spectrum. A PARAFAC2 model with three factors explains more than 99%. Figure 5a and 5b show the chromatographic loadings and the relative concentrations for factor 1 (solid), factor 2 (dashed) and, what is more important, for the background (dotted).

The two chemical analytes have been identified as acetic acid-hexyl ester (Figure 5c) and 3-hydroxy-2-butanone (Figure 5d), respectively. The background signal is most likely due to column bleeding and thus will have a specific and consistent mass spectrum within the narrow elution time window investigated and from this can be described in one PARAFAC2 factor (Figure 5e). The chromatographic loadings are perfectly defined for each sample, handling the problems of shifts within elution time profiles and across the samples, overlapping of peaks and baseline drift. The elution time loading for the background (Figure 5a) shows how the influence of the background almost remains constant throughout the different samples.

The similarity between the background mass spectrum obtained with PARAFAC2 and the mass spectrum obtained experimentally (E.I. mass spectrum obtained from [54]) is more than 99.8%. The similarities for factor 1 and factor 2 mass spectra are 95.5% and 98.9%, respectively. It has to be pointed out that the shape of the chromatographic loading for one sample (red, Figure 5a) does not have a very well defined shape. This is merely an artefact due to the way the elution profiles are presented. As can be seen in the concentration mode (red circle, Figure 5b), the concentration is approximately zero in the sample, hence, the (normalized) elution profile is not well defined.

4.3. Low intensity peaks

The next two examples are devoted to showing how PARAFAC2 can help to obtain the mass spectrum of an analyte whose signal-to-noise ratio is very low. Qualitative as well as quantitative analysis is then difficult because the mass spectrum of the analyte is highly affected by the background mass spectrum. The elution time regions between 4340-4390 scans (Figure 3c) and 5500-5700 scans (Figure 3d) reflect problems encountered with low intensity peaks. Such areas are often not studied because of the low quality of the signal and the high interference of the baseline drift.

The first situation (Figure 3c) reflects again a situation where some analytes are absent in some samples, and the peak is highly affected by the background interference, that promotes low signal-to-noise ratios. As can be expected, PARAFAC2 allows modelling the chromatographic profile even if there is no analyte (Table 2, Figure 6a). The estimated mass spectrum obtained for factor 1 identifies the peak as ethyl dec-9-enoate (Figure 6c), and the chromatographic loading reflects that there are two samples with a concentration of the analyte close to zero (Figure 6a and 6b). Once again, the similarity between obtained mass spectrum for background and experimental one is above 99.5%.

In the second situation (Figure 3d) the peak of the analyte perfectly separated from the background using two factors is obtained using PARAFAC2 (Figure 7a and 7b). In this case, the chromatographic peak has been identified as a common contaminant in a GC-MS analysis, bis-(2-ethylhexyl)phatalate (Figure 7c). The obtained mass spectrum for the background perfectly matches with the experimental one (similarity above 99.5%).

4.4. High overlapping peaks.

As evidenced above, PARAFAC2 has a remarkable ability to extract the chemical information even in case of low intense signals. Nevertheless, situations where the model cannot handle the chromatographic data have to be mentioned and studied as well.

Apart from the time of model calculation (that can be minutes to hours on current state-of-the-art personal computers) the main problem of PARAFAC2 is the need to determine the correct number of factors.

An example of the difficulty in determining the number of factors can be found in the area between scans 2100 and 2500 (Figure 3e). Apparently, there are two peaks in this elution time interval, so a PARAFAC2 model with two factors would be expected to model both peaks (maybe three if baseline is modelled as well). Nevertheless, the chromatographic loadings obtained for the model with two factors are very different from the expected (results not shown). The first factor involves both peaks; whereas the second factor seems to be an artefact of the mathematical model. This result may indicate an incorrect number of factors have been chosen and both peaks may belong to the same analyte, but in isomeric forms. Nevertheless, these peaks have been identified as isoamyl alcohol and hexanoic acid ethyl ester, respectively. The mass spectra of these analytes are indeed different.

Further analysis of the raw data between 2250 and 2350 scans indicates that the mass spectrum obtained at scan 2306 (Figure 8a) and scan 2325 (Figure 8b) for sample 1 (as an example) are different in shape. Ideally, all the mass channels have to increase in intensity in the second scan with respect to the first one. Nevertheless, there are several mass channels that decrease in intensity (channels m/z 53, 54, for example). This fact may be caused by a perfect co-elution problem of several analytes in low concentration. As mentioned above, the peak was identified as Isoamylic alcohol (3-Methyl-butanol). It is well known that isomeric forms of Methyl-butanol can co-elute [55, 56].

A PARAFAC2 model with two and three factors was fitted to the first peak, obtaining models that explained 99.5% and more than 99.9% of the total variance, respectively (Figure 9, Table 2). The model with two factors seems to indicate that, effectively, the co-elution of, at least, two isomeric forms of Methyl-butanol occurs (Figure 9a). The dashed peak was identified as Isoamylic alcohol (3-Methyl-Butanol); whereas the dotted peak was identified as 2-Methyl-Butanol. Despite the apparent success of the model, several problems still remain (red-marked samples). The elution profiles obtained for the PARAFAC2 model with three factors (Figure 9b) allow the clear identification of 3-Methyl-Butanol (dotted) and 2-Methyl-butanol (dashed), having a very well defined shape. Furthermore, a third factor has been modelled (solid). Its mass spectrum profile is highly correlated with the mass spectrum of 3-Methyl-Butanol. This may indicate that, apart from the co-elution problem, there exists other variability sources in the raw dataset that does not allow obtaining a fully satisfactory PARAFAC2 model.

5. Conclusions and remarks

PARAFAC2 has been demonstrated to be a powerful tool for solving problems derived from the experimental variability in Gas Chromatography-Mass Spectrometry (GC-MS) data. The explained variance in all models was above 99%. The obtained mass spectra matched the real mass spectra, even for the background and for analytes of very low intensity and even in situations where standard chromatographic software would not be able to provide meaningful results.

Problems such as baseline drifts, differently shaped peaks, overlapping peaks or even low signal-to-noise ratios was handled by PARAFAC2. This fact and the uniqueness property of the mass spectra allow PARAFAC2 to resolve peaks with low intensity in such a way that the mass spectral loadings can be used for qualitative purposes (identification) and that the relative concentration obtained can be used for quantitative determination of the analytes.

The technique presented here can also be used for other hyphenated separation techniques such as HPLC-UV and LC-MS, where peak shift, peak shape changes and baseline contributions are often an even bigger issue.

Acknowledgements

José Manuel Amigo wants to thank Generalitat de Catalunya for his 4-month stay fellowship BE2006-2007.

References

[1] J. Sneddon, S. Masuram, J.C. Richert, Analytical Letters, 40 (2007) 1003.
[2] P.J. Rudzki, K. Bus, H. Ksycinska, K. Kobylinska, Journal of Pharmaceutical and Biomedical Analysis, 44 (2007) 356.
[3] P.M. Medeiros, B.R.T. Simoneit, Journal of Separation Science, 30 (2007) 1516.
[4] A. Beat, B. Werner, Biological Concepts and Techniques in Toxicology, Taylor & Francis, New York, 2006.
[5] E. Eljarrat, D. Barcelo, Trac-Trends in Analytical Chemistry, 25 (2006) 421.
[6] C.Y. Hao, X.M. Zhao, P. Yang, Trac-Trends in Analytical Chemistry, 26 (2007) 569.
[7] T. Liu, M.E. Belov, N. Jaitly, W.J. Qian, R.D. Smith, Chemical Reviews, 107 (2007) 3621.
[8] G. Lubec, L. Afjehi-Sadat, Chemical Reviews, 107 (2007) 3568.
[9] M.L. Fournier, J.M. Gilmore, S.A. Martin-Brown, M.P. Washburn, Chemical Reviews, 107 (2007) 3654.
[10] L.H. Hu, M.L. Ye, X.G. Jiang, S. Feng, H.F. Zou, Analytica Chimica Acta, 598 (2007) 193.
[11] M. Katajamaa, M. Oresic, Journal of Chromatography A, 1158 (2007) 318.
[12] I. Garcia, L. Sarabia, M.C. Ortiz, J.M. Aldama, Analytica Chimica Acta, 515 (2004) 55.
[13] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Journal of Chromatography A, 805 (1998) 17.
[14] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, Journal of Chromatography A, 961 (2002) 237.
[15] V. Pravdova, B. Walczak, D.L. Massart, Analytica Chimica Acta, 456 (2002) 77.
[16] C.P. Wang, T.L. Isenhour, Analytical Chemistry, 59 (1987) 649.
[17] B. Grung, O.M. Kvalheim, Analytica Chimica Acta, 304 (1995) 57.
[18] J.W.H. Wong, C. Durante, H.M. Cartwright, Analytical Chemistry, 77 (2005) 5655.
[19] G. Tomasi, F. van den Berg, C. Andersson, Journal of Chemometrics, 18 (2004) 231.
[20] T. Skov, F. van den Berg, G. Tomasi, R. Bro, Journal of Chemometrics, 20 (2006) 484.
[21] M. Vosough, A. Salemi, Talanta, 73 (2007) 30.
[22] E. Sanchez, L.S. Ramos, B.R. Kowalski, Journal of Chromatography, 385 (1987) 151.
[23] D. Arroyo, M.C. Ortiz, L.A. Sarabia, Journal of Chromatography A, 1157 (2007) 358.
[24] E. Sanchez, B.R. Kowalski, Analytical Chemistry, 58 (1986) 496.
[25] L.S. Ramos, E. Sanchez, B.R. Kowalski, Journal of Chromatography, 385 (1987) 165.
[26] C.N. Ho, G.D. Christian, E.R. Davidson, Analytical Chemistry, 50 (1978) 1108.
[27] E. Sanchez, B.R. Kowalski, Journal of Chemometrics, 4 (1990) 29.
[28] R.A. Harshman, UCLA working papers in phonetics, 16 (1970) 1.
[29] R. Bro, Chemometrics and Intelligent Laboratory Systems, 38 (1997) 149.
[30] A.K. Smilde, Chemometrics and Intelligent Laboratory Systems, 15 (1992) 143.
[31] A. de Juan, R. Tauler, Journal of Chemometrics, 15 (2001) 749.
[32] I. Garcia, L. Sarabia, M.C. Ortiz, J.M. Aldama, Analytica Chimica Acta, 526 (2004) 139.
[33] R. Tauler, Chemometrics and Intelligent Laboratory Systems, 30 (1995) 133.
[34] R. Tauler, D. Barcelo, Trac-Trends in Analytical Chemistry, 12 (1993) 319.
[35] R. Bro, C.A. Andersson, H.A.L. Kiers, Journal of Chemometrics, 13 (1999) 295.
[36] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, Journal of Chemometrics, 13 (1999) 275.
[37] T. Skov, R. Bro, Sensors and Actuators B-Chemical, 106 (2005) 719.
[38] B.M. Wise, N.B. Gallagher, E.B. Martin, Journal of Chemometrics, 15 (2001) 285.
[39] I. Garcia, M.C. Ortiz, L. Sarabia, J.M. Aldama, Analytica Chimica Acta, 587 (2007) 222.
[40] J.M.M. Leitão, J.C.G. Esteves da Silva, Analytica Chimica Acta, 559 (2006) 1.

[41] J.M.M. Leitão, J.C.G. Esteves da Silva, Chemometrics and Intelligent Laboratory Systems, 89 (2007) 90.
[42] J.M.D. Cueva, A.V. Rossi, R.J. Poppi, Chemometrics and Intelligent Laboratory Systems, 55 (2001) 125.
[43] A. Espinosa-Mansilla, A.M. de la Pena, T.C. Goicoechea, A.C. Olivieri, Applied Spectroscopy, 58 (2004) 83.
[44] I. Ovejero-Lopez, R. Bro, W.L.P. Bredie, Food Quality and Preference, 16 (2005) 327.
[45] C.B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, R. Tauler, Chemometrics and Intelligent Laboratory Systems, 84 (2006) 9.
[46] C.B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, R. Tauler, Chemometrics and Intelligent Laboratory Systems, 83 (2006) 13.
[47] J.M.F. Ten Berge, H.A.L. Kiers, Psychometrika, 61 (1996) 123.
[48] D. Ebrahimi, J.F. Li, D.B. Hibbert, Journal of Chromatography A, 1166 (2007) 163.
[49] T. Skov, R. Bro, Analytical and Bioanalytical Chemistry, 390 (2008) 281.
[50] R. Bro, Multi-Way Analysis in the Food Industry, university of Amsterdam, Amsterdam, 1998.
[51] T. Skov, D. Balabio, R. Bro, Analytica Chimica Acta., 615 (2008) 18.
[52] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Analytical Chemistry, 73 (2001) 675.
[53] PLS-Toolbox, version 3.5, Eigenverctor Research, WA, USA.
[54] ChemStation, ChemStation and Wiley library containing Aroma compounds.
[55] A.M.T. Gonzalez, M.G. Chozas, Zeitschrift Fur Lebensmittel-Untersuchung Und-Forschung, 185 (1987) 130.
[56] L. Moio, P. Piombino, F. Addeo, Journal of Dairy Research, 67 (2000) 273.

Table 1: Main three-way methods applied in hyphenated Chromatography. Features and main drawbacks.

Method	Data structure required	Decomposition	Constraints	Main Features	Drawbacks
GRAM	Three-way array **2 samples** × elution time × spectral pattern (2 × J × K)	Based on eigenvalues	Not available	Allows prediction when test sample has unknown interferents	Only for two samples. One of them has to be a known standard. More sensitive to model errors than PARAFAC2 and MCR
DTLD	Three-way array samples × elution time × spectral pattern (I × J × K)	Based on eigenvalues	Not available	Allows prediction when test sample has unknown interferents	Makes little use of additional samples.More sensitive to model errors than PARAFAC2 and MCR
PARAFAC	Three-way array samples × elution time × spectral pattern (I × J × K)	Based on ALS	E.g. non-negativity, unimodality	Allows prediction when test sample has unknown interferents and models in cases where the above fail	Data must be trilinear
PARAFAC2	Three-way array samples × elution time × spectral pattern (I × J × K)	Based on ALS	E.g. non-negativity, unimodality	Allows prediction when test sample has unknown interferents and PARAFAC2 can handle shifted data as well as baseline drifts	Can be more sensitive to noise because profiles are estimated for each sample separately
MCR-ALS	Two-way matrix J × K For many samples, unfolded across sample dimension (IJ × K)	Approximately based on ALS	E.g. non-negativity, unimodality	Allows prediction when test sample has unknown interferents and MCR-ALS can handle shifted data as well as baseline drifts	Very good initial estimations are needed and the three-way information is lost when unfolding the array. Selectivity is needed in order to get correct results.

Table 2: Explained variance (%) for the PARAFAC2 models presented. * elution time range (scans)

N. of Factors	500-570*	2635-2695	4340-4390	5500-5700	2100-2500	2200-2360
1 Factor	96.06 (Fig. 4)	78.76	98.59	96.09	97.33	98.12
2 Factors	99.95	97.25	99.39 (Fig. 6)	99.24 (Fig. 7)	99.14	99.54 (Fig. 9a)
3 Factors	-	99.57 (Fig. 5)	-	-		99.98 (Fig. 9b)

Figure captions

Figure 1: Representation of a GC-MS chromatogram for a) one sample and for b) *K* samples.

Figure 2: An example of the application of PARAFAC2 with two factors**.**

Figure 3: GC-MS dataset of 24 samples of wine. Five situations have been enhanced. (a) Shift in retention time, (b) Overlapping of peaks, (c-d) two cases of low intensity of the peaks and (e) multipeak analysis.

Figure 4: PARAFAC2 results for the area between 500 and 570 scans.

a) Chromatographic loadings, b) concentration loading, c) Comparison between PARAFAC2 mass spectrum loading (upper) and real mass spectrum (bottom) for acetic acid ethyl ester.

Figure 5: PARAFAC2 results for the area between 2365 and 2395 scans.

a) Chromatographic loadings and b) concentration loading (background, dotted; factor 1, solid; factor 2, dashed). Comparison between PARAFAC2 mass spectrum loading (upper) and E.I. mass spectrum (bottom) for c) factor 1, acetic acid hexyl ester, d) factor 2, 3-hydroxy-2-butanone and e) background.

Figure 6: PARAFAC2 results for the area between 4340 and 4390 scans.

a) Chromatographic loadings and b) concentration loading (background, dotted; factor 1, solid). c) Calculated mass spectra for factor 1, ethyl-dec-9-enoate.

Figure 7: PARAFAC2 results for the area between 5500 and 5700 scans.

a) Chromatographic loadings and b) concentration loading (background, dotted; factor 1, solid). c) Calculated mass spectra for factor 1, bis-(2-ethylhexyl)phatalate.

Figure 8: Mass spectra obtained for sample 1 at a) 2306 and b) 2325 time scans.

Figure 9: PARAFAC2 models of a) two and b) three factor of the area compressed between 2200 and 2360 time scans.

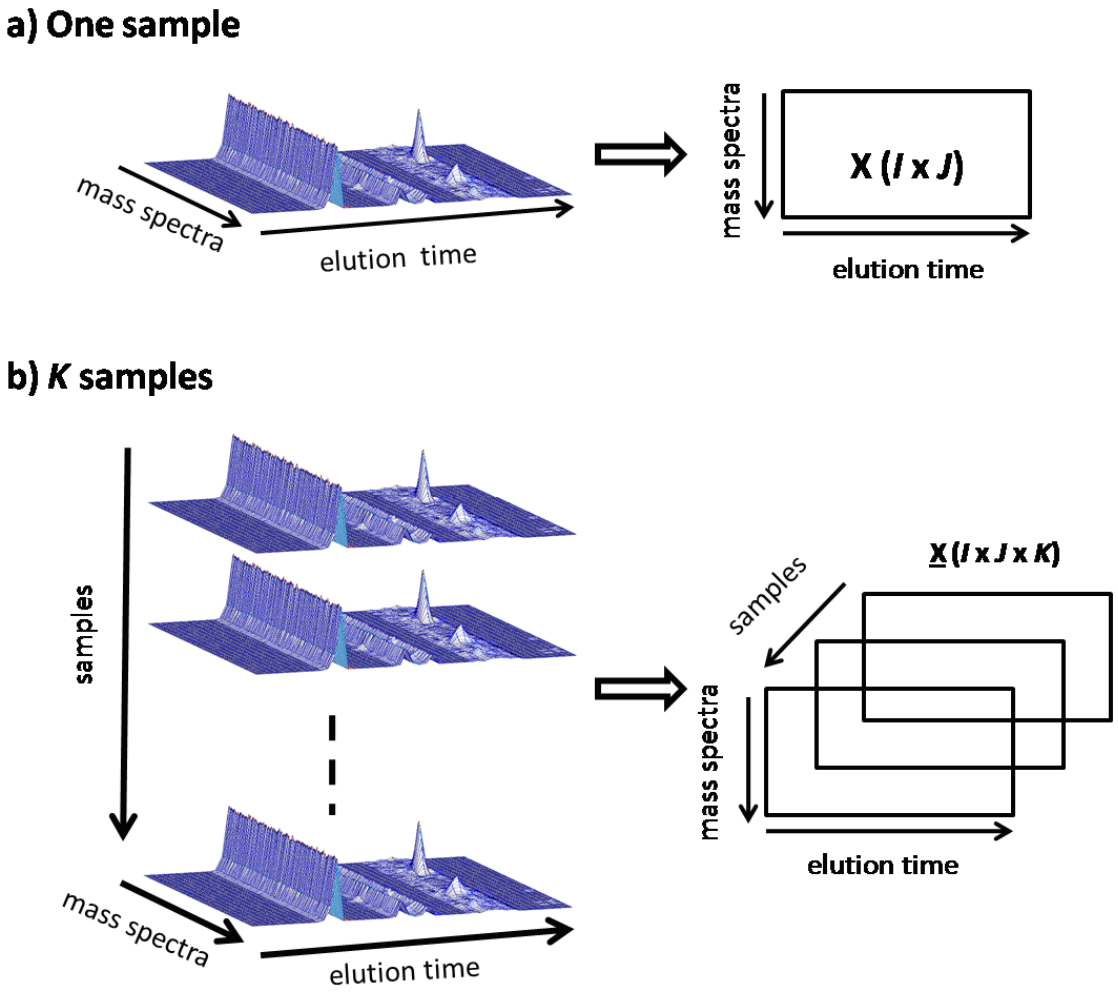

a) One sample
mass spectra
elution time
mass spectra
X (I x J)
elution time
b) K samples
samples
mass spectra
elution time
X (I x J x K)
samples
mass spectra
elution time

FIGURE 1

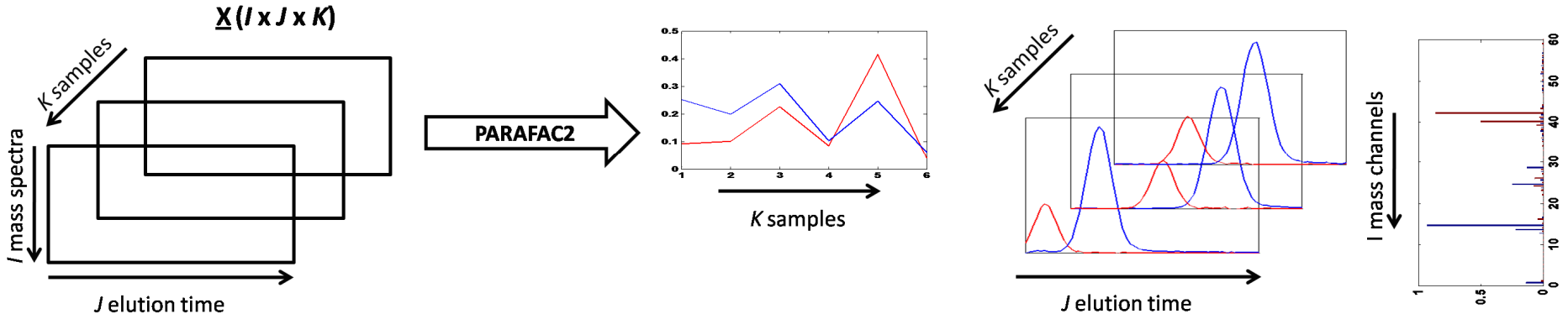

X (I x J x K)
K samples
I mass spectra
J elution time
PARAFAC2
K samples
K samples
J elution time
I mass channels

FIGURE 2

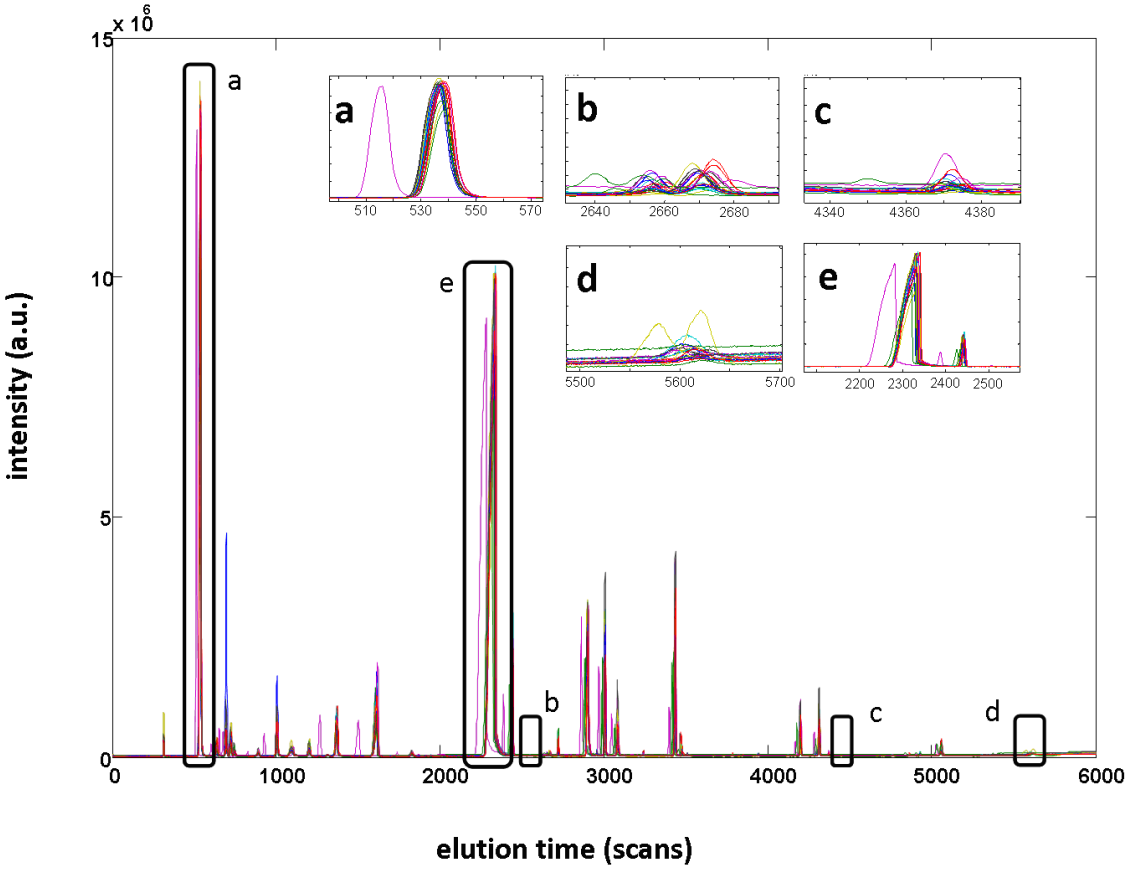

x 10^6
15
10
5
0
intensity (a.u.)
0
1000
2000
3000
4000
5000
6000
elution time (scans)
a
b
c
d
e
510
530
550
570
2640
2660
2680
4340
4360
4380
5500
5600
5700
2200
2300
2400
2500

FIGURE 3

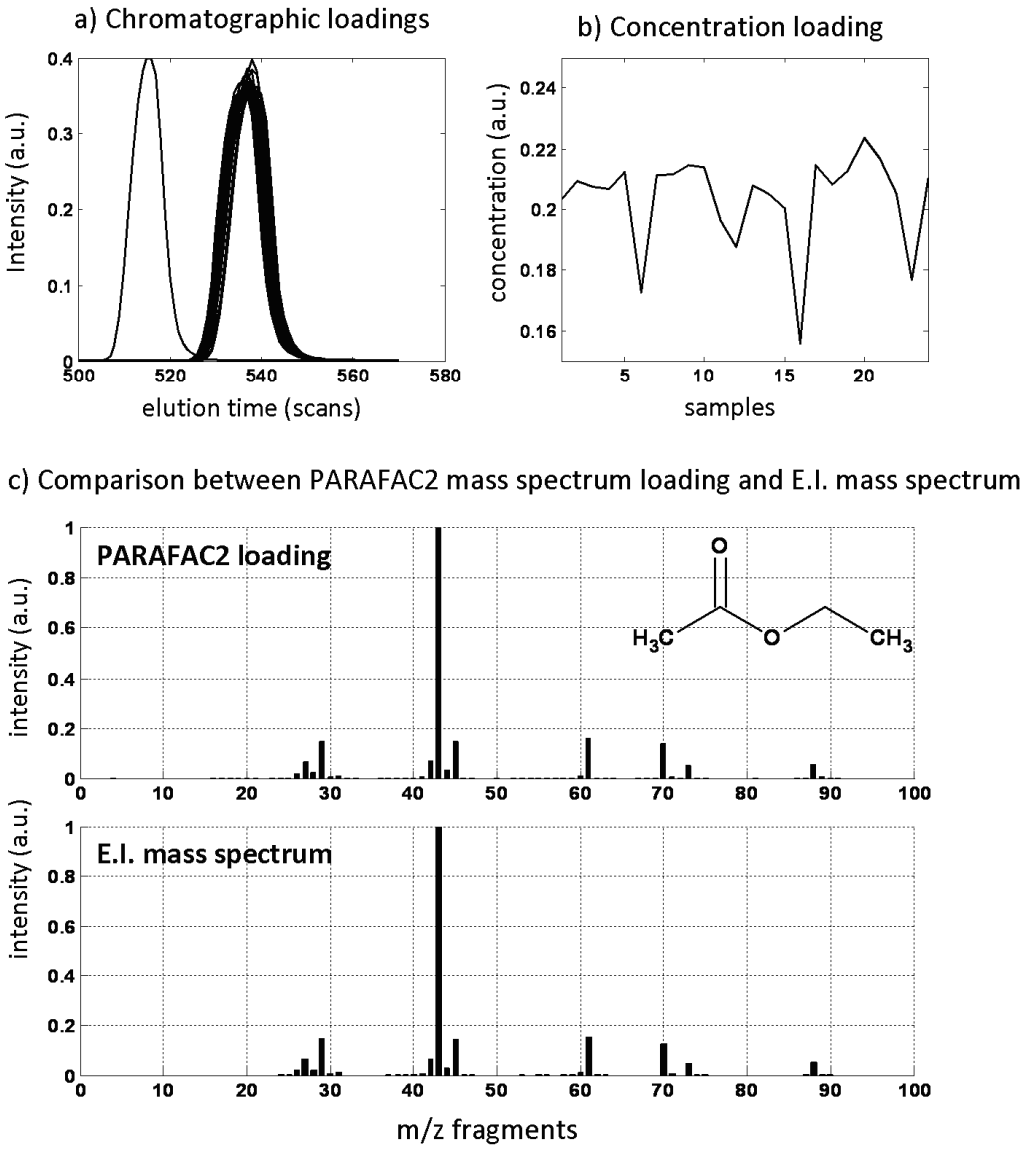

a) Chromatographic loadings
Intensity (a.u.)
elution time (scans)
b) Concentration loading
concentration (a.u.)
samples
c) Comparison between PARAFAC2 mass spectrum loading and E.I. mass spectrum
PARAFAC2 loading
intensity (a.u.)
E.I. mass spectrum
intensity (a.u.)
m/z fragments

FIGURE 4

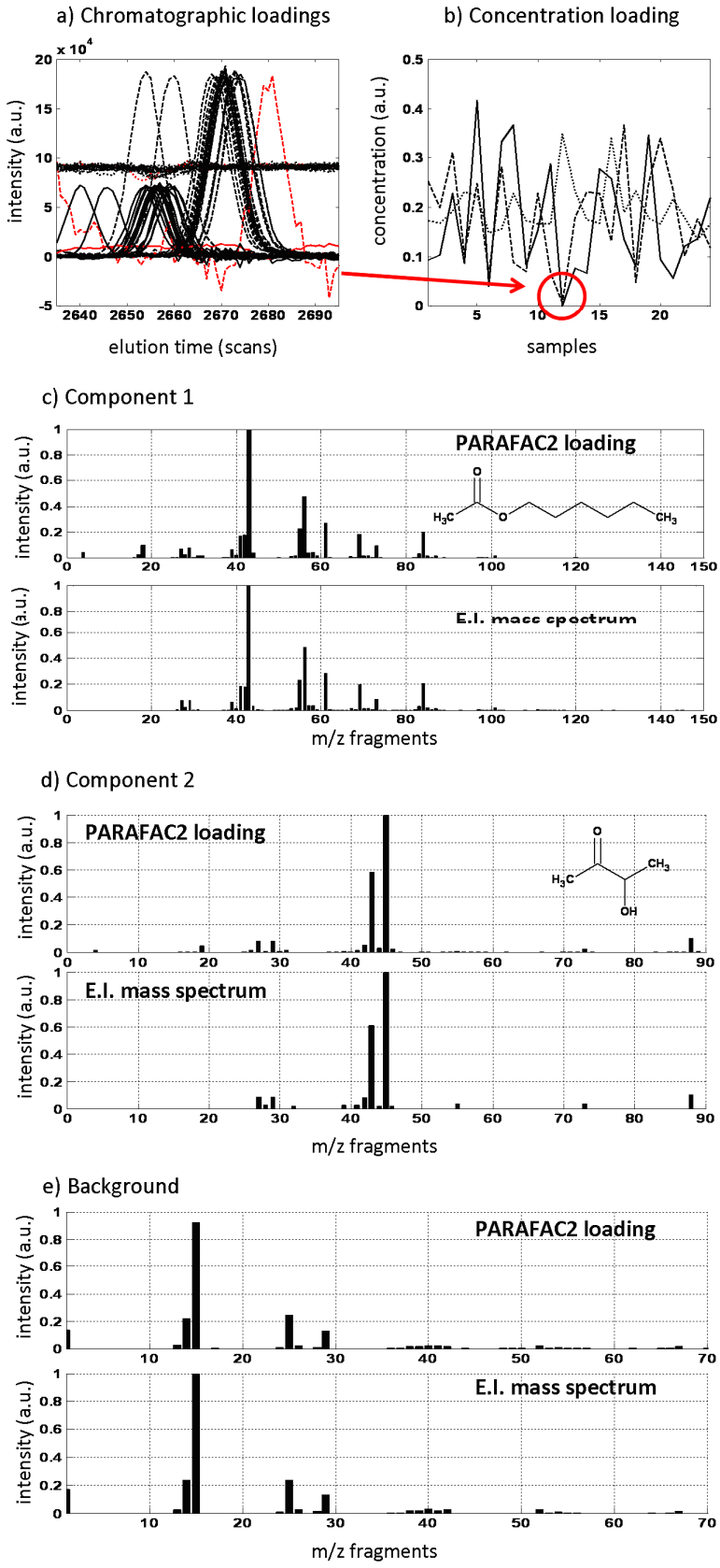

a) Chromatographic loadings
b) Concentration loading
x 10^4
intensity (a.u.)
elution time (scans)
concentration (a.u.)
samples
c) Component 1
PARAFAC2 loading
E.I. mass spectrum
m/z fragments
d) Component 2
PARAFAC2 loading
E.I. mass spectrum
m/z fragments
e) Background
PARAFAC2 loading
E.I. mass spectrum
m/z fragments

FIGURE 5

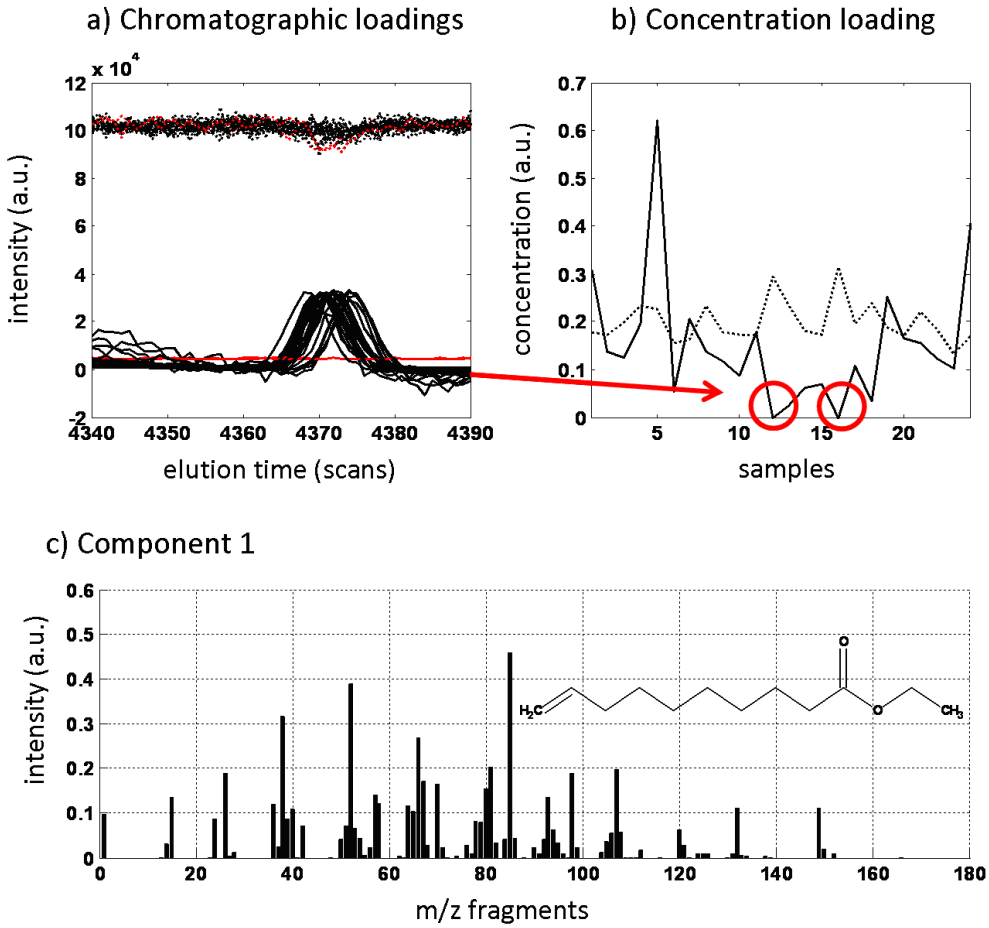

a) Chromatographic loadings
b) Concentration loading
x 10^4
intensity (a.u.)
elution time (scans)
concentration (a.u.)
samples
c) Component 1
intensity (a.u.)
m/z fragments
H2C
O
O
CH3

FIGURE 6

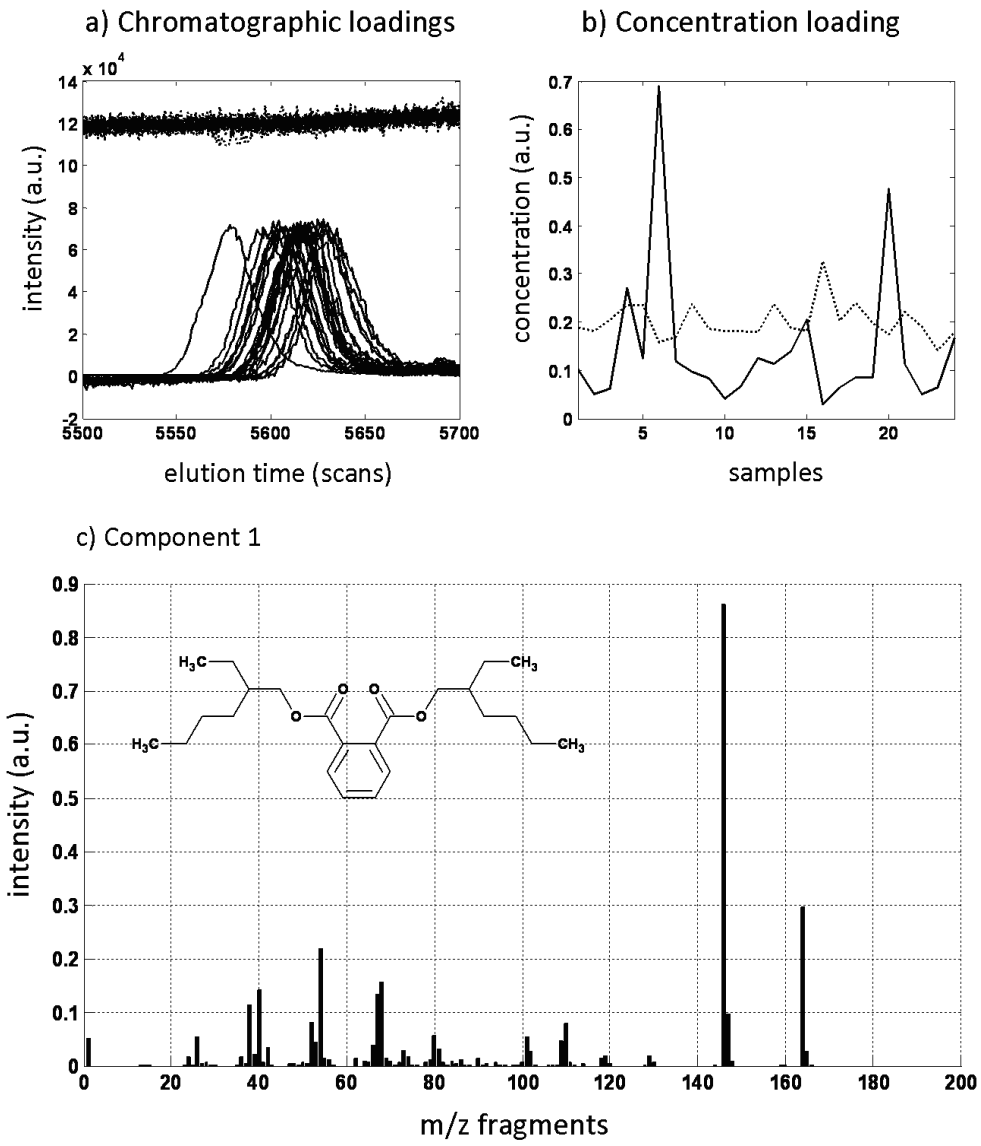

a) Chromatographic loadings
x 10^4
intensity (a.u.)
elution time (scans)
b) Concentration loading
concentration (a.u.)
samples
c) Component 1
intensity (a.u.)
m/z fragments

FIGURE 7

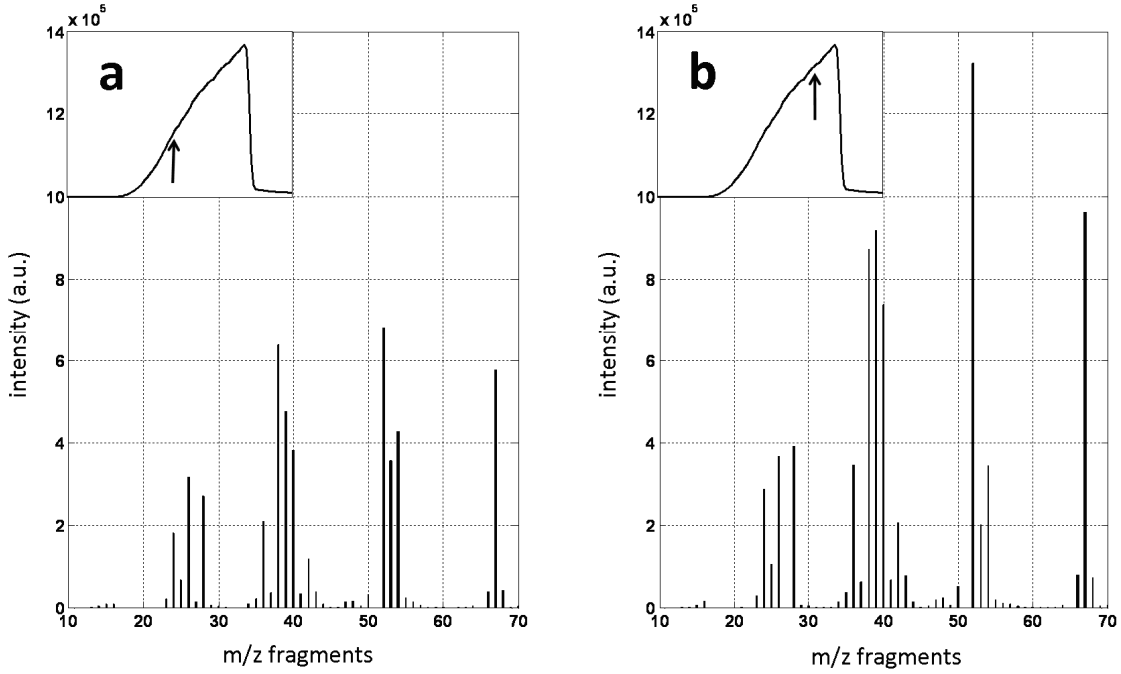

a
b
x 10^5
intensity (a.u.)
m/z fragments
0
2
4
6
8
10
12
14
10
20
30
40
50
60
70

FIGURE 8

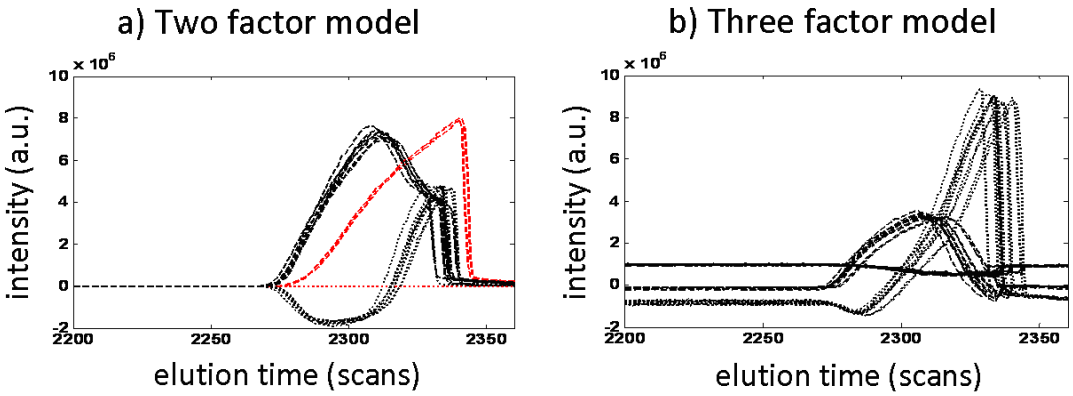
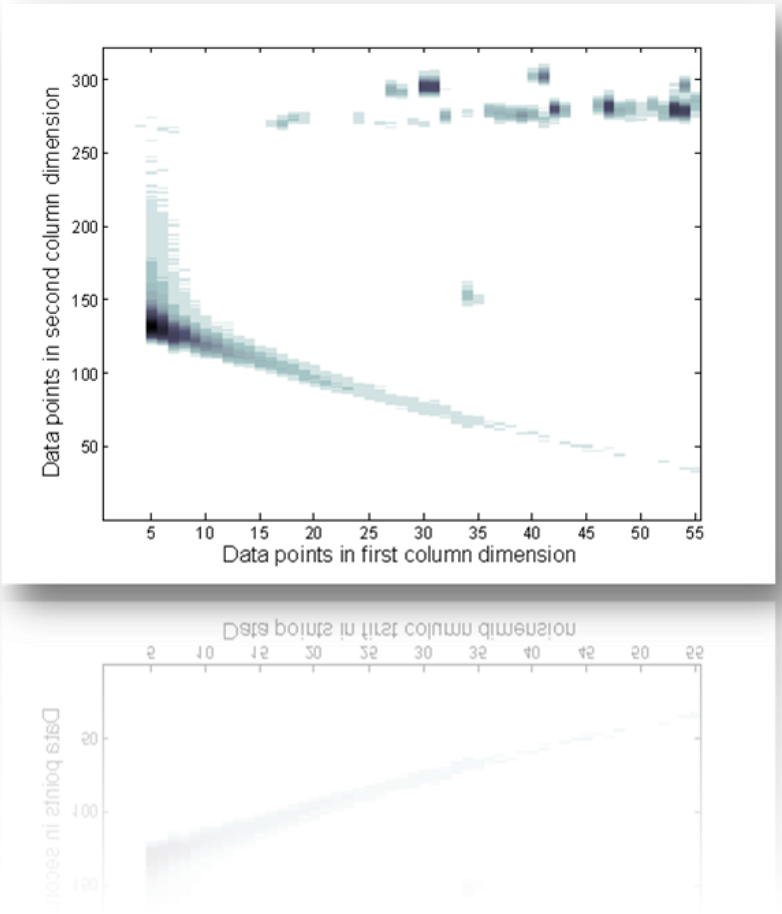

a) Two factor model
b) Three factor model
x 10^6
intensity (a.u.)
elution time (scans)
2200
2250
2300
2350

FIGURE 9

PAPER V

Skov, T., Hoggard, J.C., Bro, R., and Synovec, R.E. **2008**. Handling Retention Time Shifts in GC×GC-TOFMS Data using Shift Correction and Modeling. *In preparation*.



Handling Retention Time Shifts in GC×GC-TOFMS Data using Shift Correction and Modeling

Thomas Skov[1*], Jamin C. Hoggard[2], Rasmus Bro[1] and Robert E. Synovec[2]

[1]Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.

[2]Department of Chemistry, Box 351700, University of Washington, Seattle, Washington 98195-1700

*corresponding author, email: thsk@life.ku.dk

Prepared for submission to Journal of Chromatography

Keywords: GC×GC-TOFMS, retention time shift, PARAFAC, PARAFAC2

Abstract

The use of PARAFAC for modeling GC×GC-TOFMS peaks (i.e., mathematical resolution for identification and quantification) is well documented. This success of PARAFAC is due to the trilinear structure of these data under ideal, or sufficiently close to ideal, chromatographic conditions. However, using temperature programming to cope with the general elution problem, deviations from trilinearity are more likely to be seen for the following three cases: (1) compounds (i.e., analytes) severely broadened on the first column hence defined by many modulation periods, (2) analytes with a very high retention factor on the second column and likely wrapped around in that dimension, or (3) with fast temperature program rates. This deviation from trilinearity is seen as retention time-shifted peak profiles in subsequent modulation periods (first column fractions) and this will hamper the use of PARAFAC. In this report, a relaxed yet powerful version of PARAFAC, known as PARAFAC2 has been applied to handle this shift within the model step by allowing generation of individual peak profiles in subsequent first column fractions. An alternative approach was also studied, utilizing a standard retention time shift correction to restore the data trilinearity structure followed by implementing PARAFAC. These two approaches are compared when identifying and quantifying a known analyte (bromobenzene) over a large concentration series where a certain shift is simulated in the successive first column fractions. Finally, the methods are applied to real chromatographic data showing severely shifted peak profiles. The pros and cons of the presented approaches are discussed in relation to the model parameters, the signal-to-noise ratio and the degree of shift.

1. Introduction

Comprehensive two-dimensional gas chromatography (GC×GC) is becoming an established technique for the analysis of complex mixtures [1-20]. When it is combined with time-of-flight mass spectrometry (TOFMS), the result is a very powerful instrument (GC×GC-TOFMS) [21-31]. With GC×GC-TOFMS, for each sample run a cube of data is obtained, rich with chemical information. While it is a powerful technique, and much has been done in the area of development of chemometric software to analyze such data [5-7,15-20,27,30,31-35] there are still considerable challenges that remain in the area of improving data analysis methods. Data from the GC×GC-TOFMS measurement of a single chemical compound (i.e., analyte) has what is referred to as a trilinear structure dependent upon the two chromatographic separation axes and the mass spectral axis. An ideal trilinear structure would result in a constant retention time of a given analyte on the second column for each slice (per column one fractions that are driven by the instrument modulation period). However, deviations from trilinearity can be observed under non-ideal chromatographic conditions (e.g., analyte overloading and/or severely broadened peaks in the first dimension), or during rapidly changing separation conditions on the first dimension as during a fast temperature program. For a single sample run, some deviation from trilinearity can be expected due to an applied temperature program to deal with the general elution problem. In most practical implementations of GC×GC-TOFMS, the two GC columns are normally temperature programmed at about the same rate, or nearly so; thus, the temperature increase per time for the first column will simultaneously correlate to a temperature increase on the second column. The result of this can be observed as a change in second column retention time for a specific analyte found in succeeding first column fractions (i.e., for each subsequent modulation period). An example of this deviation from trilinearity is visualized in Figure 1.

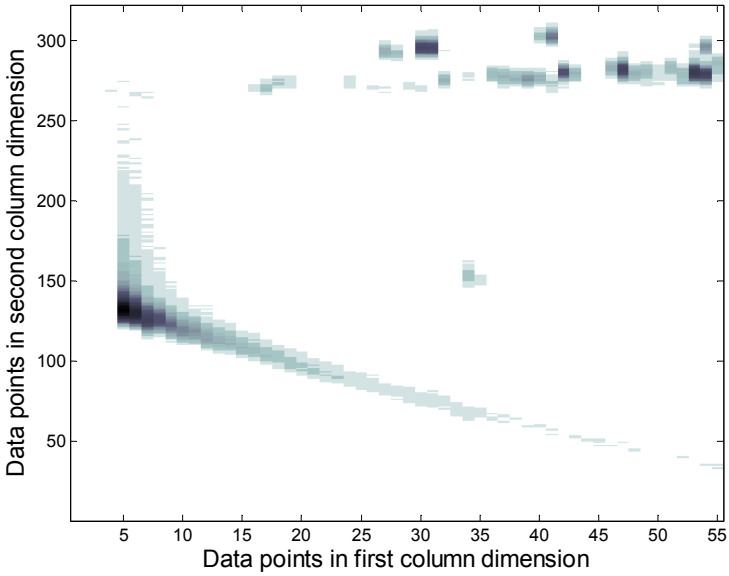


Figure 1. Illustration of the shifted peak profile position observed when applying a temperature program on the first column in a GC×GC instrument where the two columns are housed together. The main peak shown is a dimethyl phosphite peak (50 ppm) spiked into a kerosene sample. The first and second column retention times are given as data points or scans. Approximately 161 scans (mass spectra) are acquired per second, which gives a second column retention time span of 2 s (modulation period). Only the contribution from the ion fragment m/z = 80 is illustrated to enhance the dimethyl phosphite peak.

In earlier studies it has been shown that the low-rank trilinear PARAFAC model is able to model a single analyte in GC×GC-TOFMS data adequately if the data is indeed low-rank trilinear [21,22,32-36]. With PARAFAC it is assumed that each first column fraction (data point in Figure 1) containing the analyte (in different amounts) can be characterized by an estimated common profile for the second column chromatographic profile and a common mass spectral profile. If this is the case, a unique solution with an estimate of each chromatographic profile and mass spectrum can be obtained. However, this assumption is not valid with sufficiently shifted peak positions in the second column dimension over the first column fractions.

To set the scene, this report starts by theoretically explaining the shift in second column retention time using the experimental chromatographic conditions behind the GC×GC-TOFMS analysis shown in Figure 1. After that, two different approaches to handle this shifted behavior will be discussed in detail. The first approach is to apply a pre-processing method that corrects for the shift and thus, restore the trilinearity. This is done using a rather simple integer retention time shift correction. For the shift corrected data, PARAFAC [37,38] will be used to find qualitative and quantitative information. The second approach is to apply PARAFAC2 [39,40], an extended and less restricted version of PARAFAC, designed to handle shift located within one specific dimension and across one of the other dimensions. PARAFAC2 allows one unique chromatographic profile for each first column fraction to be calculated under certain constraints to maintain the uniqueness of the model. This property has spread the use of the PARAFAC2 model to such diverse data structures as LC-Fluorescence [40], sensor based data [41], batch chemical processes [42], HPLC-DAD [43], time intensity profiles [44] and GC-MS [45], but it has not yet been implemented and evaluated for GC×GC-TOFMS data.

The two approaches will be validated using different model-derived characteristics and external knowledge of the analyte investigated. Firstly, a novel method that compares the estimated mass spectral loading with the analyte mass spectrum found from a commercial library by a so-called match value will be used [33]. Secondly, the total sum of all elements in the three-way array of the restored analyte peak will be used as a measure of the relative concentration of the analyte. This will be calculated for the same analyte over a large concentration series to evaluate the performance of the two multiway methods at different signal-to-noise ratios.

2. Theory

2.1 Changes in second column retention time

Deviations from tri-linearity are seen in GC×GC-TOFMS data because the retention time (also denoted elution time) of an analyte on a given column depends on, among other things, the temperature of the column and stationary phase. As temperature increases, analytes are less retained on the stationary phase, spend more time in the mobile phase, and thus, elute earlier. Temperature programs are commonly used in GC and GC×GC to deal with the general elution problem. The temperature of the second column is generally increased throughout a run along with the temperature of the first column, partly because of instrumental restrictions (the second column commonly being housed in an oven within the oven the first column is in), but also when using two independent ovens, to prevent analytes from being retained too short or long on the second column as would happen if the temperature of the

second column was kept constant. The run-to-run temperature increase on the second column causes analyte retention times on the second column to decrease from slice to slice taken from the first column (assuming a constant flow rate), and thus induces a shift in second column retention times across a peak [16]. A theoretical relation between retention time and temperature for constant flow rate is given as follows:

$$\ln k' = \ln\left(\frac{t_{R,2} - t_{0,2}}{t_{0,2}}\right) = \frac{\Delta H}{R \cdot T} + C \quad (1)$$

Where k' is the retention factor, $t_{0,2}$ is the dead time along the second column, $t_{R,2}$ is the retention time of the analyte along the second column, ΔH is the enthalpy of vaporization of the analyte, R is the gas constant, T is the temperature of the column (also the stationary phase and carrier gas), and C is a value that depends on the interaction of the analyte with the stationary phase, the phase volume ratio, as well as other effects (such as entropy) that can be assumed to be constant under the conditions of interest.

Solving for $t_{R,2}$ gives:

$$t_{R,2} = t_{0,2} e^{\frac{\Delta H}{R \cdot T} + C} + t_{0,2} \quad (2)$$

The enthalpy of vaporization changes only slowly with temperature, and the dead time of the column system is constant with constant flow, so these terms can be assumed to be constant across the temperature range in which an analyte will elute. Equation (2) can be used to calculate the theoretical slice-to-slice shift for a peak (i.e., from one modulation period to the next), as will be demonstrated next using data from the peak shown in Figure 1.

The temperature of the second column when the first slice of the peak elutes can be calculated using the first column retention time and temperature program for the second column (given in the Experimental section); in this case, the temperature was ~ 100 ºC. Next, the dead time of the second column, $t_{0,2}$, is either measured using a dead time marker or calculated using a suitable tool. Using the LECO GC×GC-TOFMS Column Calculator (LECO Corp., St. Joseph, MI), the dead time is 1.36 s for the separation conditions used for Figure 1. The first slice of the peak has an apparent retention time of 0.82 s; this is less than the dead time, so the peak must have wrapped around the 2 s modulation period and eluted in a subsequent modulation. The actual retention time could be any combination of (2.82 + 2n) seconds, where n is any nonnegative integer, but only small values of n are probable because the peak is not likely to be retained much longer than several seconds on the second column. From previous isothermal high speed GC experiments (constructing a van't Hoff plot) the enthalpy of vaporization for dimethyl phosphite is known to be 37.7 kJ at ~100 ºC. Using this information, C can be calculated from Equation (1) for any of the likely retention times, and expected slice-to-slice shifts across the peak can be calculated using Equation (2). In fact, if there is no ambiguity in the number of times an analyte has wrapped around the second separation dimension (e.g., $t_{0,2}$ and $t_{R,2}$ are known for several slices),

Equation (1) can be used to construct a van't Hoff plot and determine ΔH of a compound using corresponding analyte data from a single GC×GC chromatogram — however, the error in ΔH will likely be larger than that obtained from the traditional van't Hoff plot construction using several isothermal GC chromatograms because the temperature range spanned is relatively small (unless the peak is unusually broad on the first column, or if the temperature program rate is high).

Table 1. Theoretical predictions of slice-to-slice shifting along the second column for the first slice of the dimethyl phosphite peak shown in Figure 1. $\Delta t_{R,2}$ actually decreases slightly across the peak (as temperature increases), but the change is small enough to be ignored under these conditions.

Parameters	Less retained	Medium retained	Longer retained
Observed $t_{R,2}$ (s)[#]	1.50	2.82	4.82
k'	0.103	1.07	2.54
C	–14.4	–12.1	–11.2
$\Delta t_{R,2}$ (s per slice)[#]	–0.00151	–0.0158	–0.0373

[#]This number can be converted to scans or data points by multiplying with 161.29 scans/s.

The average second column retention time shift, $\Delta t_{R,2}$, of the dimethyl phosphite peak observed across three replicates of the data shown in Figure 1 is –0.0153 s (equivalent to –2.47 data points). This number is very close to the theoretical slice-to-slice shift calculated given a second column retention time of 2.82 s (center column of Table 1). The left column of Table 1 gives results for a compound that is less retained, and shows that the slice-to-slice shifting is expected to be much smaller in less retained compounds than in longer retained compounds, as seen in the middle and right columns. Although not seen in Table 1, the modulation period is another factor affecting the amount of shifting. At a given temperature program rate, peaks in runs using longer modulation periods exhibit greater changes in retention time from one modulation slice to the next because the time between modulations is longer and thus the change in temperature from slice-to-slice is greater. The sampling rate also affects the amount of observed shifting.

Another effect that can cause changes in second column retention times across a peak are changes in flow rate caused by the temperature program. The flow rate decreases with increasing temperature when using a constant head pressure in a GC×GC method, which in turn contributes to increasing the second column retention times across a peak. For runs set for a constant flow rate, instruments usually adjust head pressure based on calculations involving the column dimensions to try to maintain a constant flow rate. Differences in actual or effective column dimensions or operating parameters from those used in calculations can lead to changes in flow rate over the course of temperature programmed run and resulting changes in retention time.

Aside from the causes given above, small deviations in modulation timing can cause slight changes in second column retention time from slice to slice. Most modulators available today, however, are well-timed and synchronized to injection, so this effect is small compared to those mentioned above. Of the different sources of retention time deviation, those induced by temperature programs are the most common and the most severe and will thus be studied in this report, although the methods herein should

be equally applicable to any shifting situation. It is important to note that the effects that cause deviations from trilinearity are more pronounced on broader peaks along the first separation dimension. The width of peaks depends on, among other things, the interaction between the stationary phase and the analyte. Thus, a sample separated using a stationary phase ideal for one class of compounds may yield much broader peaks for another class of compounds within the same sample, as is the case with the spiked kerosene samples presented below (Figure 2).

In Figure 2 four different temperature programs have been applied for the analyte dimethyl phosphite in a kerosene oil sample. Only the signal intensity from the mass channel m/z = 80 has been plotted for clarity.

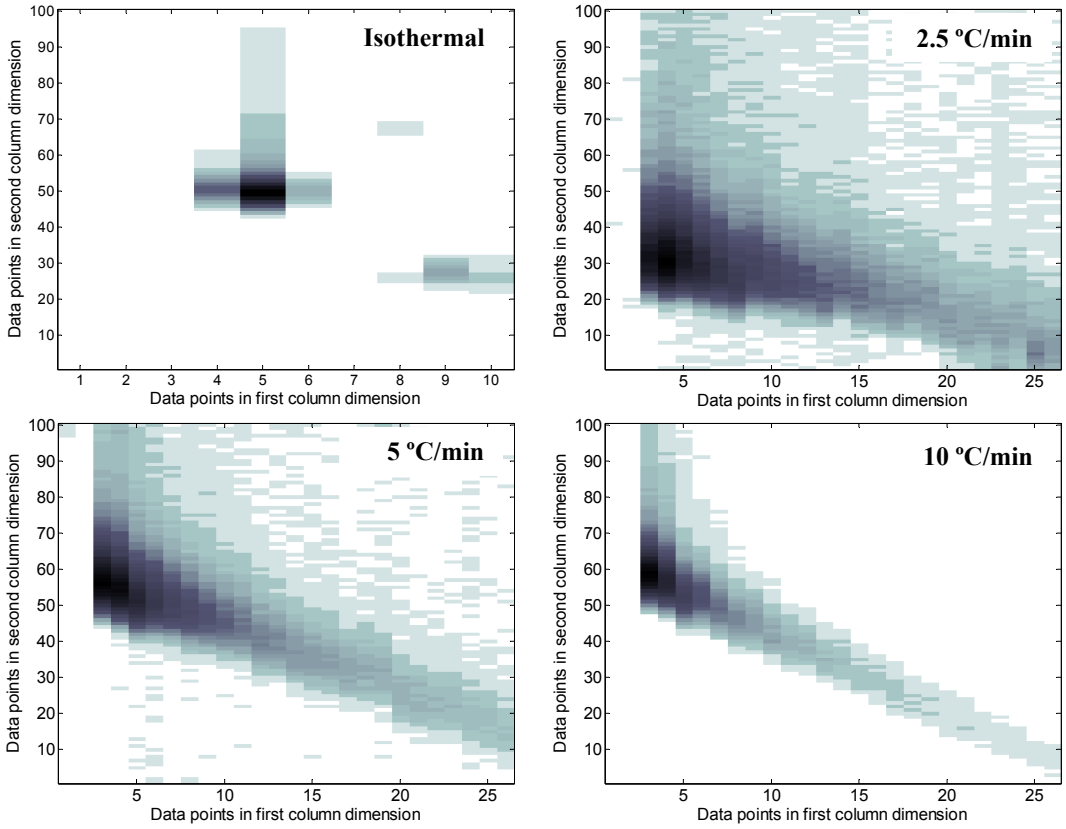


Figure 2. Illustration of the shifted first column slices when applying different temperature programs on the first column for the dimethyl phosphite peak (50 ppm) in kerosene samples. Notice that the same length of the second column dimension (same number of data points and thus time) has been used to highlight the degree of shift for the three temperature settings. Also notice that a reduced second column elution time has been plotted here compared to Figure 1. Only the contribution from the ion fragment *m*/*z* = 80 is illustrated to enhance the shift of the dimethyl phosphite peak. The 10 ºC/min run is further visualized in Figure 3.

In Figure 3 the temperature program of 10 ºC/min has been plotted in a different way to highlight the chromatographic profile of the individual first column slices.

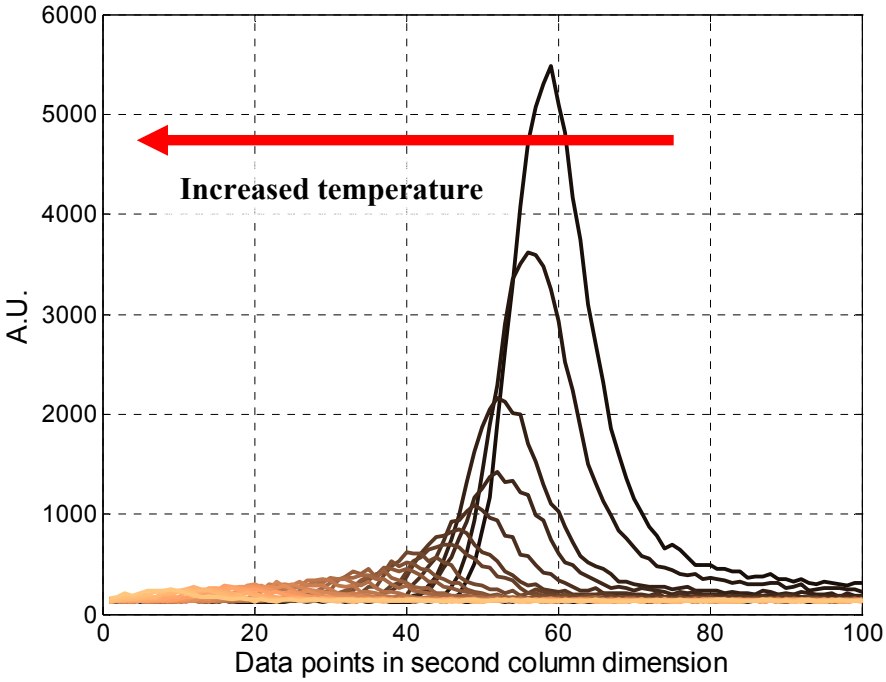


Figure 3. The temperature program of 10 ºC/min plotted to show the chromatographic nature of the data. As seen also from Figure 1 the later eluting fractions (first column slices) has a shorter elution time due to the high temperature program used.

Figure 2 shows that applying a higher temperature program shortens the retention time of the analyte on the second column and as such makes the use of strictly trilinear models less appropriate. Retention time shifted data have been extensively described in the literature for one dimensional chromatographic data (e.g., GC-FID, TIC from GC-MS), but shift observed in two-dimensional chromatographic data (e.g., GC×GC) has not been extensively evaluated, with only a few reports [46,47]. Indeed shift corrections in GC×GC-TOFMS data that uses a range of the m/z collected (not just the TIC) has not been reported until now. Here we put forward two methods for dealing with this phenomenon; either to correct the shifted data (restoration of the trilinearity via retention time alignment) or to use a more advanced multiway model suitable for dealing with shifted behavior in the data (PARAFAC2).

The three-way structure of GC×GC-TOFMS data will be exploited using low-rank multiway models on local peak regions having one or only a very few peaks, taking one sample at a time. As each experiment has shift introduced in the second column dimension, the analysis of more than one sample at a time is unfeasible as some shift in peak position is expected across the samples as well. Besides this, no multiway model for handling this double dimensional shift has been presented yet.

2.2 Multiway models

Trilinear models such as PARAFAC have been widely used for the analysis of individual peaks or multiple peaks for both GC-MS and GC×GC-TOFMS [21,22,32-38,48,49] data.

For GC-MS, data from several samples are stacked providing a three-way array whereas individual samples are modeled for GC×GC-TOFMS data. PARAllel FACtor Analysis (PARAFAC) [37,38] is a trilinear model that decomposes the data array into three loadings matrices as shown in Figure 4.

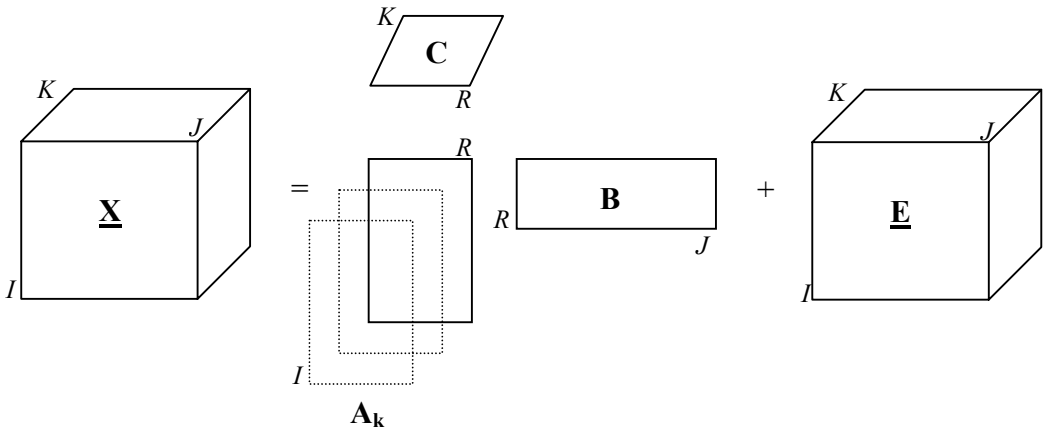


Figure 4. Graphical description of the PARAFAC/2 models. Data: GC×GC-TOFMS. Mode **A**: Second column elution time; mode **B**: Mass channels (m/z) and mode **C**: First column elution time. For the PARAFAC model only one **A** matrix of size $I \times R$ exists (Equation 3). The data is arranged according with the algorithm used for the PARAFAC2 model [see section 3.3]. Here the shift is located within the first mode (second column elution time) and across the third mode (first column fractions - i.e. first column elution time).

The matrix formulation of the PARAFAC model [37,38] and PARAFAC2 model [39,40] are shown below.

$$\text{PARAFAC:} \quad \mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}} + \mathbf{E}_k \qquad k = 1,..,K \tag{3}$$

$$\text{PARAFAC2:} \quad \mathbf{X}_k = \mathbf{A}_k\mathbf{D}_k\mathbf{B}^{\mathrm{T}} + \mathbf{E}_k \qquad k = 1,..,K \tag{4}$$

where $\mathbf{X}_k$ is the kth frontal slab of the three-way array and $\mathbf{D}_k$ is a diagonal matrix holding the kth row of **C** in its diagonal. **A**/$\mathbf{A}_k$, **B,** and **C** are parameters to be estimated and $\mathbf{E}_k$ residuals. The major difference between the two models is that PARAFAC2 allows the loading matrix for the second column mode to be different for the k first column slices ($\mathbf{A}_k$).

PARAFAC is suitable for complex data with low-rank trilinear structure, like GC×GC-TOFMS data signals, where chromatographic profiles of individual analytes are not completely resolved by the

instrument. With PARAFAC, the overlapped peaks can often be mathematically resolved if they are sufficiently different in the second column profile and mass spectral profile. This mathematical resolution, or deconvolution, does not require peak shape assumptions or completely selective mass channels [32]. The first is only valid as long as the most concentrated fraction introduced on the second column does not result in a peak shape change, e.g., peak tailing due to overload of the column. This would cause a deviation from trilinearity; a fundamental problem in chromatographic analysis and a problem that is difficult to solve even with highly advanced pre-processing techniques (alignment, peak shape modeling, etc.).

This fundamental problem can, together with the peak shifts illustrated in Figure 1, be solved and handled by a more advanced multiway model called PARAFAC2. Unlike PARAFAC, the PARAFAC2 model relaxes the strict trilinearity by allowing profiles to be estimated in one mode (second column) for each occasion in the other mode (first column mode in Figure 4) [39,40]. In the case shown in Figure 2 (and more easily on Figure 3), PARAFAC2 could be applied introducing twenty six sets of estimated elution time profiles; one for each of the first column fractions (i.e., modulation period or slices), thereby being able to describe different peak shapes and peak positions in different fractions runs. The principles of the PARAFAC2 model have also been shown in Figure 4, where several loadings matrices are estimated; one for each k slab ($\mathbf{A}_k$). PARAFAC2 has to estimate more elements in the model step (uses more degrees of freedom) and thus, will be affected by noise (low intense signals) in a different way than PARAFAC.

Both PARAFAC models will provide *one* best-fit solution uniquely under mild conditions if the proper number of components has been determined. This means that the solution cannot be rotated without a loss of fit and, most importantly, that meaningful chemical model parameters (e.g., pure spectra and elution time profiles) are found when the proper number of components are used. For more on properties of the uniqueness of the two models, the reader is referred to [37-40].

2.2.1 Model performance

There are several ways of determining the proper number of factors for multiway models and four important ones are 1) split-half analysis [50], 2) judging residuals, 3) core consistency diagnostic [51], and 4) compare with external knowledge of the data being modeled [33]. Here the latter three approaches will be pursued focusing mainly on the last one, which has been proven to be very efficient for GC×GC-TOFMS data with known a analyte [33].

The factor describing the analyte of interest is found from the similarity between the estimated mass spectral loadings and a library spectrum [33,52]. When this value (a so-called match value) exceeds a threshold (here 750 is used as proposed by [33]) the factor will be assigned to describe the analyte of interest. To find the largest match value the proper number of components in the model must be estimated. This was done using the so called two-factor degeneracy where the analyte information has been split over more than one factor. This results in two components in the same model with highly correlated loadings profiles – a clear indication of overfit. The model with the fewer factors is then chosen as the model describing the data and the factor holding the analyte used for further diagnostics and evaluation. If no splitting is observed, the model holding the largest match value above 750 is

chosen among the valid models. The match value range is from 0 to 1000 where 1000 indicates a perfect match over all mass channels. The same approach is used with success for PARAFAC2 to validate if the analyte of interest is being modeled.

After confirmation that the right analyte has been found, modeled properly and that the information is kept in one specific factor, the model is evaluated further looking at residuals and core consistency values. This is done to make sure that the match value obtained is indeed due to the known analyte being modeled and not due to an uncertain estimation of factors.

2.2.2 Quantification of analyte

The quantification of the analytes in GC×GC-TOFMS can be done in many ways. Here the approach based on summing the elements of the outer product of the loadings of the component describing the analyte of the PARAFAC/2 models will be used. Mathematically this sum, *S*, can be written as follows:

$$S_{PARAFAC} = \sum \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} \tag{5}$$

Where **a**, **b**, and **c** are the three loading vectors representing the second column, the mass spectral and the first column loadings, respectively, of the component describing the analyte of interest, respectively in accordance with the structure of the PARAFAC models illustrated in Figure 4.

For the PARAFAC2 model the quantification is essentially the same, but instead of one common **a** vector, *k* loading profiles are found and thus *k* outer products are calculated, one for each corresponding first column slice, and summed, as follows:

$$S_{PARAFAC2} = \sum_{k} \sum \mathbf{a}_k \otimes \mathbf{b} \otimes c_k \tag{6}$$

Another approach for finding the analyte concentration that is worth a short comment is the use of the score value (loading of the first dimension). If data are rearranged so that the first mode holds the first column elution time, then the scores of that mode would be a measure of the magnitude of the unique second column and mass spectral loadings. Summing these score values would give a similar peak sum as the one described above. First column slices containing the peak would then have a much higher score value than slices only holding noise and/or background. This is only valid for PARAFAC, where the second and third loading vectors are normalized to a length of one and thus, the first column scores will be a measure of the content of these loadings. For PARAFAC2 the first column slices must be in the last mode to follow the convention for the PARAFAC2 algorithm used. With this structure, the mass spectral loading (mode **B**) and the first column loading (mode **C**) are normalized to a length of one and thus, the first column loading cannot be used in the same way as for PARAFAC. However, normalizing each second column loading vector ($\mathbf{A}_k$) to one, and rescaling the loading of the first column elution time so that the reconstructed data ($\mathbf{X}_k$) remains the same, then the loading of the first column elution time will describe the amount as explained above for PARAFAC. This means that the score value can be compared across different peaks and sample models.

2.2.3 Model constraints

The solution from multiway models can be constrained in order to include external knowledge (e.g., from the analytical technique) in the fitting procedure and to help the model find the chemical phenomena in the data. For GC×GC-TOFMS the natural constraints would be to force the model to have non-negative values in the estimated loadings as this would fit the chromatographic profiles and the mass spectra. The use of non-negativity constraints in all three modes has previously been found to introduce an offset in the chromatographic loadings [33]. A method to handle this offset was reported for the PARAFAC model using a rather simple baseline correction procedure. Presuming a Gaussian shape of the analyte peak, two chromatographic regions at ±3 standard deviations or more away from the peak maximum in each of the chromatographic dimensions of the PARAFAC model were selected. Baselines were constructed for the two chromatographic dimensions in each model by taking the average of all the points in each region and then fitting a line through the two resulting points. The resulting baselines were subtracted from the loadings to give what will be referred to as bias-corrected loadings. This approach works well for the PARAFAC parameters.

For PARAFAC2, non-negativity was applied only in the first column and mass spectra mode due to what is possible with the current PARAFAC2 algorithms. A small baseline offset was also observed here, but due to the negative values found in the loadings for the shifted mode this offset was corrected in the reconstructed peak region. This was done by finding all slices holding no peak structure and then subtracting the average of these slices from the reconstructed peak signal (a way to find these slices is presented in section 3.3.2). The use of a similar baseline correction method for GC^3 (GC×GC×GC) [53] showed a very good similarity between the peak sum from baseline corrected data and PARAFAC modeled data.

3. Experimental

3.1 Sample preparation

3.1.1 Dimethyl Phosphite

A spike solution containing dimethyl phosphite in acetone at a concentration of 2.03×10^{-3} g/mL was prepared in a 5 mL volumetric flask. Kerosene was added to a 5 mL volumetric flask containing 0.0968 g (123 μL) of the spike solution up to the mark, such that the concentration of dimethyl phosphite was 5.00×10^{-5} g/mL, or ~ 50 ppm.

3.1.2 Bromobenzene

Samples containing bromobenzene were prepared by serial dilution with hexane in the following concentrations, in g/mL: 1.01×10^{-7}, 3.18×10^{-8}, 1.03×10^{-8}, 3.19×10^{-9}, 1.02×10^{-9}, 3.17×10^{-10} and 1.01×10^{-10}. Although samples were made in and contained within tightly capped vials having tetrafluoroethylene (TFE) seals, samples were kept in a deep freezer when storage was necessary to minimize evaporation.

The seven concentrations of bromobenzene were all injected in four replicates. These data were used due to the wide range of concentrations (different signal-to-noise ratios) so limits in shift correction method and multiway models to model analytes of very different concentrations can be evaluated. However, the peak shifts in these data were not significant enough to illustrate the effectiveness of the shift correction methods and the multiway models, so an artificial peak shift was introduced as explained in section 3.3.3.

In Table 2 the selected and used mass channels for the two analytes evaluated in this study are shown. In both cases top ten intense mass channels were selected.

Table 2. Mass channels (m/z) selected for bromobenzene and dimethyl phosphite.

#	Bromobenzene	Dimethyl Phosphite
1	77	80
2	156	79
3	158	47
4	51	95
5	50	49
6	38	109
7	78	65
8	74	110
9	75	48
10	157	77

3.2 Instrumentation

Samples were analyzed using a LECO Pegasus 4D GC × GC-TOFMS System (LECO Corp., St. Joseph, MI). A 20 m × 250 μm column with 0.5 μm 5% diphenyl/95% dimethylpolysiloxane film (DB-5; J&W Scientific/Agilent Technologies) was used as the first column, and a 2 m × 180 μm column with 0.2 μm

trifluoropropylmethyl polysiloxane film (Rtx-200; Restek, Bellefonte, PA, USA) was used as the second column. Ultra high purity helium was used as the carrier gas.

3.2.1 Dimethyl Phosphite

For the kerosene samples, the inlet temperature was set at 250 °C and the flow rate was set to a constant 1.0 mL/min. A 50:1 split was used on the injection of 1.0 μL of sample. The transfer line between the GC oven and mass spectrometer was set 235 °C. The mass spectrometer was set to collect *m*/*z* 40 to 269 and bin 31 of the 5000 transient spectra/s to give ~161.29 spectra/s. The ion source was set to 200 °C and electron energy of –70 V. The mass spectrometer detector voltage was set to –1700 V.

Four different temperature programs were used for the kerosene samples: an isothermal program maintaining a temperature of 150 °C in the first column oven and 160 °C in the second column oven for 15.00 minutes, then ramping to 230 °C and 235 °C (for the first and second column ovens respectively) at 25 °C/min and holding these temperatures for 10.00 min; a program ramping at 2.5 °C/min (after beginning at 60 °C and 70 °C for half a minute) to 230 °C and 235 °C on the first and second column ovens respectively, then holding these temperatures for 5.00 min; a program ramping at 5.0 °C/min using the same temperatures and hold times as the 2.5 °C/min program; and a program ramping at 10 °C/min, again using the same temperatures and hold times as the 2.5 °C/min program. For all temperature programs, the modulation period was 2.0 seconds and the modulator was set to maintain a temperature 20 °C above that of the first column oven.

3.2.2 Bromobenzene

For the bromobenzene samples, a split-less injection of 1.0 μL of sample from the auto sampler delivered the sample to the inlet at 200 °C. The separation was performed with a constant flow rate of 0.80 mL/min. A temperature program was used, starting at 60 °C and 70 °C for 0.25 min, then ramping at 8 °C/min up to 110 °C and 120 °C for the first and second column ovens respectively. The modulator was kept at 40 °C above the first column oven temperature and set to modulate one time each second. The transfer line temperature was held at 280 °C. Ion source temperature in the mass spectrometer was set at 200 °C, and an electron energy of –70 V was used. An *m*/*z* range from 5 to 160 was collected in every transient spectrum, and 50 of these transient spectra collected at 5 kHz were binned to give 100 spectra/s. The mass spectrometer detector voltage was set to –2000 V.

3.3 Data analysis

Collected data were exported from LECO software as comma-separated value (.CSV) files and imported into MATLAB 7.3 (The MathWorks, Natick, MA). The PARAFAC and PARAFAC2 algorithm from the PLS toolbox version 4.0 (Eigenvector, Manson, WA) [54] were used.

3.3.1 Shift correction method

A shift correction is important if trilinear models such as PARAFAC models should be used in a parsimonious way. Here the shift correction is based on calculating the correlation coefficient between adjacent first column slices.

Principles

This method moves slices one point at a time in a small region around the peak of interest. The shift correction leading to a maximum correlation coefficient is used as the shift needed for each first column elution time slice provided that the difference between the maximum and the minimum correlation coefficient is above a certain threshold. The method calculates correlation coefficients for each individual mass channel, *j*, for all adjacent slices and uses a mask to explore if shifts are considered valid or not, according to:

$$\text{Mask}_j = (\text{CorrCoef}_{\max} - \text{CorrCoef}_{\min}) \geq \text{threshold} \quad (7)$$

If the difference is above the threshold value, the mask is set to one. For each comparison of adjacent slices the correlation coefficient in each mass channel that contains a valid shift is summed and divided by the sum of the quantity of the correlation coefficient times the shift needed to give the corresponding valid highest correlation coefficient (valid shifts). The outcome is an overall shift correction for each first column slice based on a valid contribution from each individual mass channel. As seen the shift correction method include a difference between correlations and not just the absolute correlation between adjacent slices. This is a rather strict approach making sure that artifacts, interference and noise will not guide the shift correction procedure. Thus, only the presence of an analyte will result in a valid shift correction. However, this rigidity means that very low-signal-to-noise analytes can be difficult to align, but is has been found that this fits well with the cautiousness that must be taken in low concentration samples.

If the analyte of interest is eluting in the outer regions of the whole second column dimension some problems will arise with this method. One way to deal with this problem is to insert values in the regions. These elements must be similar to the background and/or noise. To solve this problem, a method to estimate background and/or noise elements has been put forward in section 3.3.2.

3.3.2 Extending the second column for proper shift correction

If the peak of interest is eluting halfway with respect to the total second column elution time then enough space around this peak will be available when aligning. On the other hand, if the peak is placed close to one border (min or max elution time) and a severe shift is present, then the slices to be shift corrected will contain missing values. In this case new values must be inserted to extend the second column dimension to allow for proper shift correction and these values should be as close to the original background and/or noise values as possible. Two approaches will be used depending upon the peak shift observed, the position of the peak and interfering analytes.

Extending using artificial elements resembling the background/noise

An interfering peak or the peak of interest itself will provide structure over adjacent slices. A measure for this structure is the correlation between adjacent first column slices, as peaks are always spread over more slices (see e.g., Figure 2). If the correlation is found between all adjacent slices for all mass channels, the slices containing no peak structure in any mass channel can be found. These slices then hold elements that are free of analyte(s) and thus, only describe the background and/or noise. It was found that a correlation of 0.7 resulted in a proper selection of slices without any structural information

(i.e., peaks) and these slices were then used to calculate mean and standard deviation of the background and/or noise (mass-channel-wise). This approach was carried out for both first and second column elution time slices and if non-structure holding slices were found in both dimensions, then these were all selected.

For each mass channel, normally distributed random values (with the found mean and standard deviation) were inserted to extend the second column dimension. In this way the mass spectrum of the background is preserved and the magnitude (and variation) of the inserted elements will be of the same size as the original background elements.

3.3.3 Artificial shifts

The efficiency of the integer shift correction method and the multiway methods relative to each other is mainly affected by one thing, the signal-to-noise ratio. A lower signal-to-noise ratio results in a decreased correlation between adjacent slices and makes it more difficult for the method to differentiate between analyte and noise and/or interferences.

To test both shift correction method and multiway models, artificially shifted data were created using a simple integer movement in one direction as illustrated in Figure 5. Artificial shifts introduced to all replicates of the seven concentrations of bromobenzene will be used for the study of the effect of shift correction methods and multiway models. One can argue that introducing an artificial integer shift and then correcting with an integer shift correction method might be superfluous. However, the success of the presented multiway models and the shift correction method are affected by the signal-to-noise ratio and thus, the likely found difficulties for both alignment method and modeling when lowering the concentration will be easier to evaluate.

Here an artificial shift of two data points was used. This number was selected based on knowledge about the range of expected shifts over different temperature programs and for different analytes. As shown in Figures 1, 2 and 5 and in Table 1, a slice-to-slice shift in the range of zero to six data points/scans or more could be observed under different experimental conditions for different analytes. The standard deviations between peak maximum of adjacent slices (within-run second column deviations) were also investigated. It was found that the average of the standard deviations between slices (in many samples) was approximately 0.45 scans. This indicates that some uncertainty will be found even for samples run under isothermal conditions. However, as the uncertainty was so small (a deviation of less than one scan cannot be corrected by the integer shift correction method) it was not included in the artificial shifts.

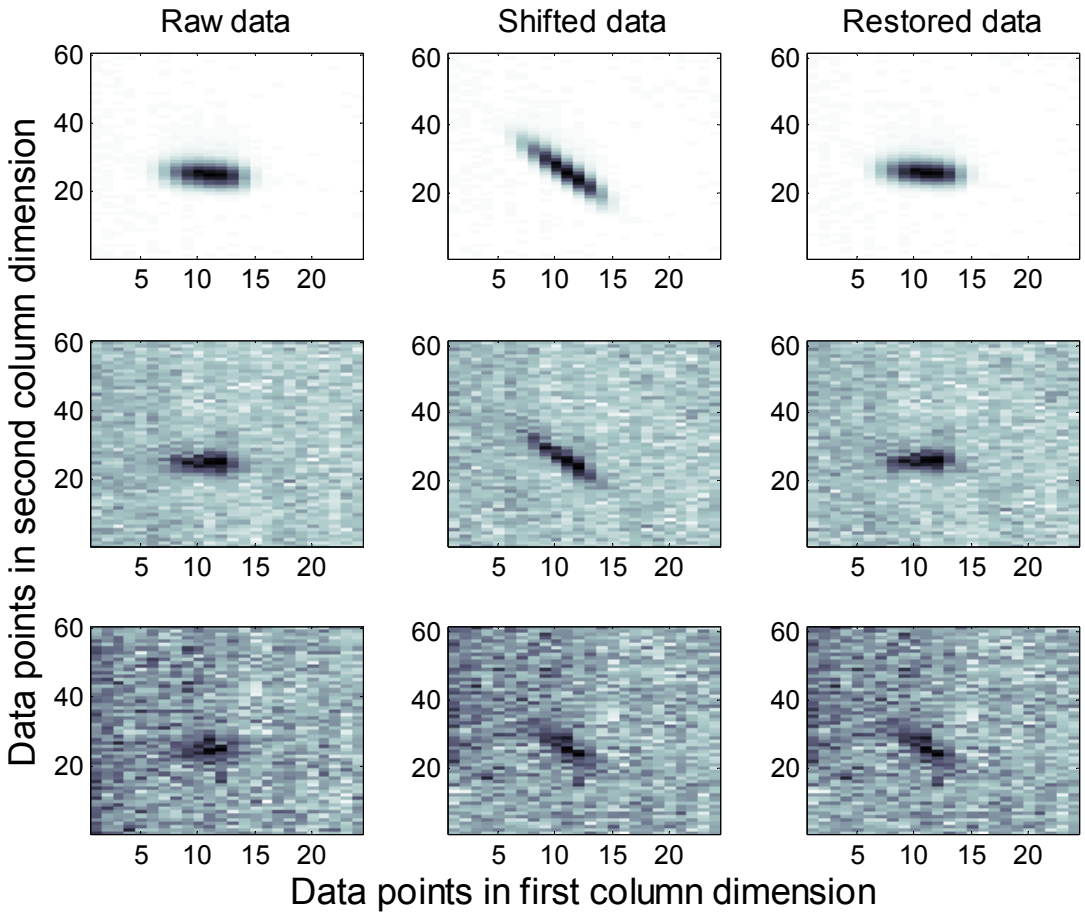


Figure 5. Examples of raw data (first column), artificially shifted data (second column) (see section 3.3.4) and restored data using integer based method (third column). TOP: one replicate of bromobenzene in a concentration of 1.01×10^{-7}, MIDDLE: one replicate of bromobenzene in a concentration of 3.19×10^{-9} and BOTTOM: one replicate of bromobenzene in a concentration of 1.02×10^{-9}. Only the contribution from the ion fragment $m/z = 77$ is illustrated to enhance the slice-to-slice peak shift of bromobenzene.

As mentioned the shift correction method applied is rather conservative and thus does not correct a potential shift unless verified over several mass channels. This is also seen in Figure 5, where the lowest concentration of bromobenzene is not restored correctly. This is the borderline where it is difficult to see which method to use: use shift correction followed by a low-rank trilinear model (PARAFAC) or to use a model handling shifted profiles within the model (PARAFAC2).

3.3.4 Number of factors in PARAFAC models

As discussed in a previous report [33], the number of factors for the optimal PARAFAC model can be determined based on the match value. For all models calculated on the seven different concentrations of bromobenzene, this approach was applied for models with 1 to 5 factors. The maximum number of factors was determined based on initial investigations of the capability of the model to find the bromobenzene peak even at very low concentrations. For the lowest concentrations it was not possible to find and quantify the bromobenzene peak using this approach even with more factors included in the models (up to 10 factors was tested). For the other concentrations, the bromobenzene peak was

described in models containing from 2 to 5 factors. The general trend was that the lower the concentration of bromobenzene, the more factors were needed and smaller match values were achieved. This can be explained by the decreased signal-to-noise ratio [33].

For the dimethyl phosphite samples the number of factors was the same as described above, but for the samples run at isothermal conditions the number of factors was increased to 7 for both PARAFAC and PARAFAC2 due to more interfering analytes in the peak region investigated (see Figure 2).

For both analytes, models having different numbers of factors provided similar match values before splitting was observed. In these situations model parameters such as core consistency, residuals and loadings were further investigated to select the proper model. If two models were found equally good at describing the analyte of interest (i.e., similar match value, loadings and peak sum, etc.), the model of lowest complexity was selected.

4. Results and discussion

In the following, all PARAFAC and PARAFAC2 models have been optimized according to the principles explained in the Experimental section. The results will be presented in the following chronological order:

1. Pre-processing of data – is it needed?
2. Evaluation of model performances depending on degree of shift, shift correction method applied and signal-to-noise ratios.
3. Application to 'real' chromatographic data.

The PARAFAC and PARAFAC2 models will be used to evaluate the individual steps according to the diagram shown in Figure 6.

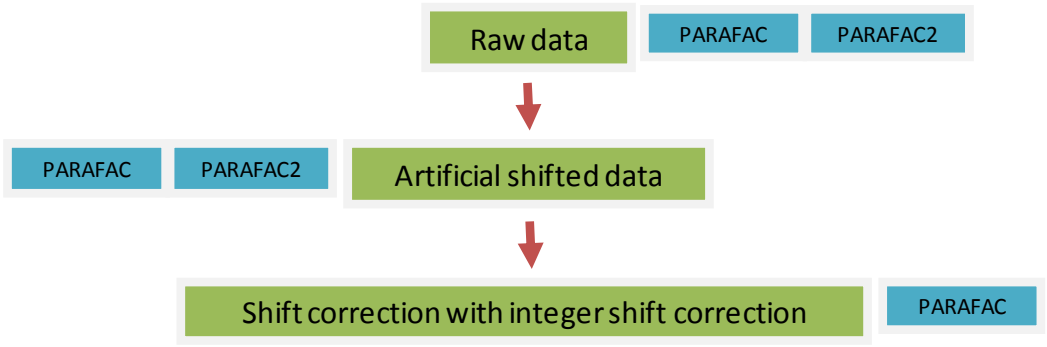


Figure 6. Diagram of how the individual data arrangements (raw, shifted, corrected etc.) were handled and which models were used in each step.

4.1 Pre-processing of data

Before doing any shift correction the use of the strict trilinear PARAFAC model for shifted data was tested. In Figure 7 the shifted data and the modeled data using PARAFAC and PARAFAC2 for one replicate in the highest bromobenzene concentration are visualized.

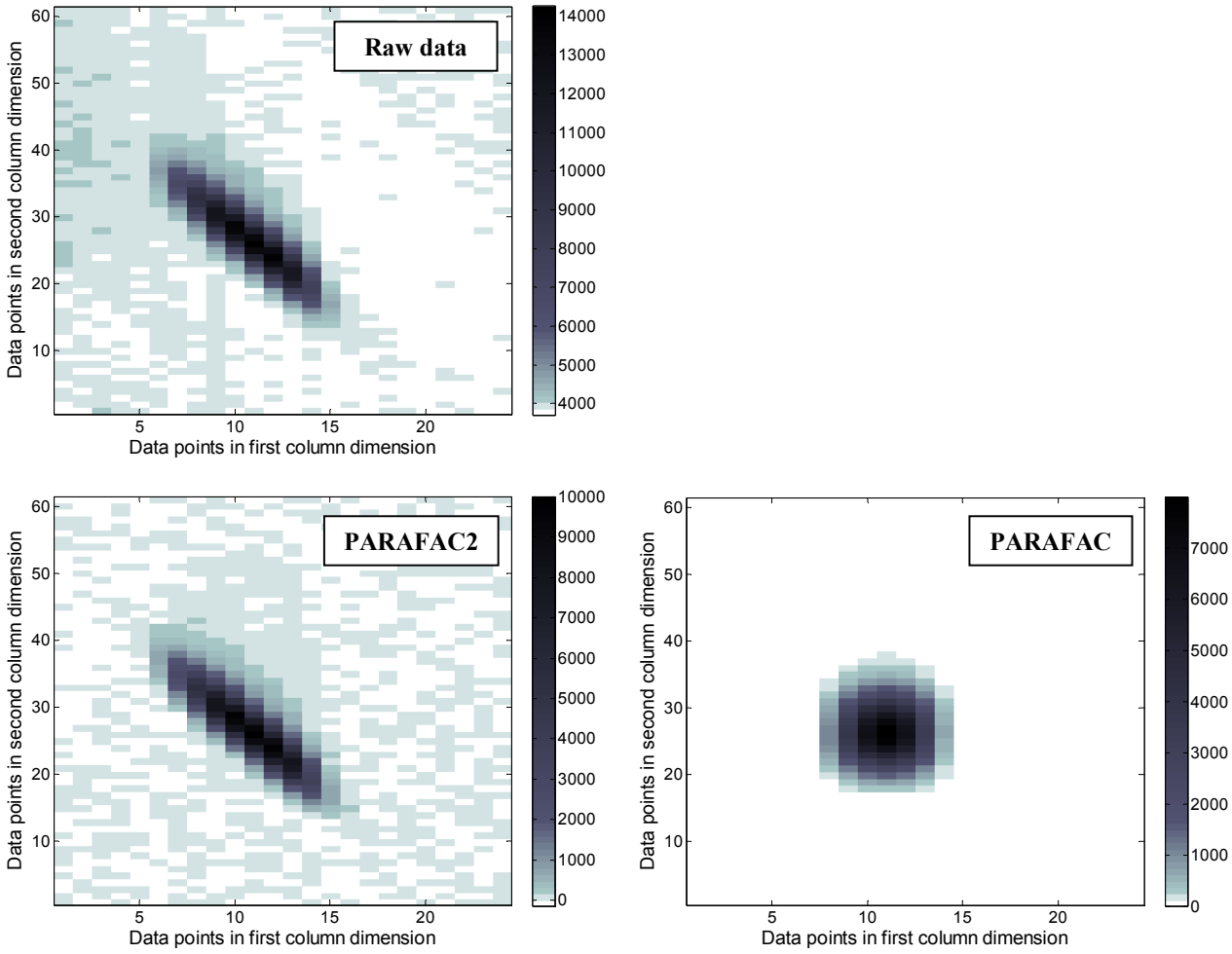


Figure 7. Visualization of PARAFAC and PARAFAC2 modeling of artificial shifted data (one replicate of the bromobenzene peak in highest concentration). Notice the scale of the contour plot. TOP: Raw data (peak + baseline), BOTTOM LEFT: PARAFAC2 (peak correctly modeled) and RIGHT: PARAFAC (peak wrongly modeled).

Figure 7 shows, that PARAFAC, as expected, cannot optimally model a peak when too severe shift (deviation from trilinearity) is originally present in data. PARAFAC2 models the peak almost perfectly and from a two factor model where one factor holds the peak and the other factor the baseline, the original raw data can be reconstructed. The assumption of common loading profiles in the PARAFAC model is also seen and this results in a poor peak description – here indicated with a change in original peak profile and significant lowered peak maximum intensity (i.e., incorrect peak sum). This small example clearly demonstrates that some kind of pre-processing retention time shift correction is needed if PARAFAC is to be used to model shifted data. One could argue that PARAFAC2 then is the model of choice for this type of data. However, the more complex model using more degrees of freedom is more affected by noise. So, with PARAFAC2, one is capable of modeling real chromatographic data and then presenting the potential shift in a visual and simple way for easy understanding for non-experts of multiway models.

4.2 Shift correction method

The shift correction has been applied to the sample shown in Figure 1 and the corrections of the dimethyl phosphite peak run with a temperature program of 10 ºC/min can be seen in Figure 8 for different threshold levels (see Equation 7).

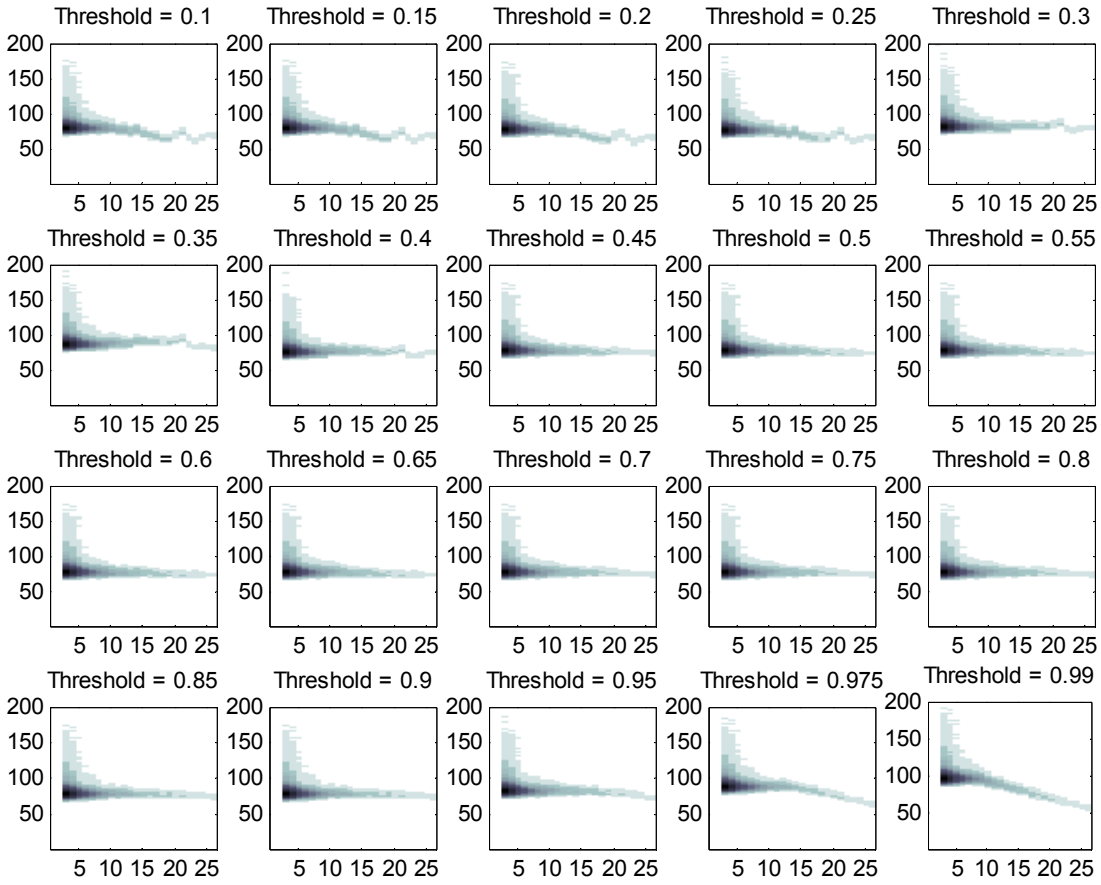


Figure 8. Illustration of a restored dimethyl phosphite peak (program of 10 ºC /min) using different threshold values. Only the contribution from the fragment m/z = 80 is illustrated to ignore small interfering analytes.

Figure 8 shows that the method can correct for the shift selecting the proper threshold value. Visually it is seen that the method is affected by a correct choice of threshold and the value must be between 0.45 and 0.90 for correction of the shift.

In Figure 9 the peak sums for the shift corrected data shown in Figure 8 are presented.

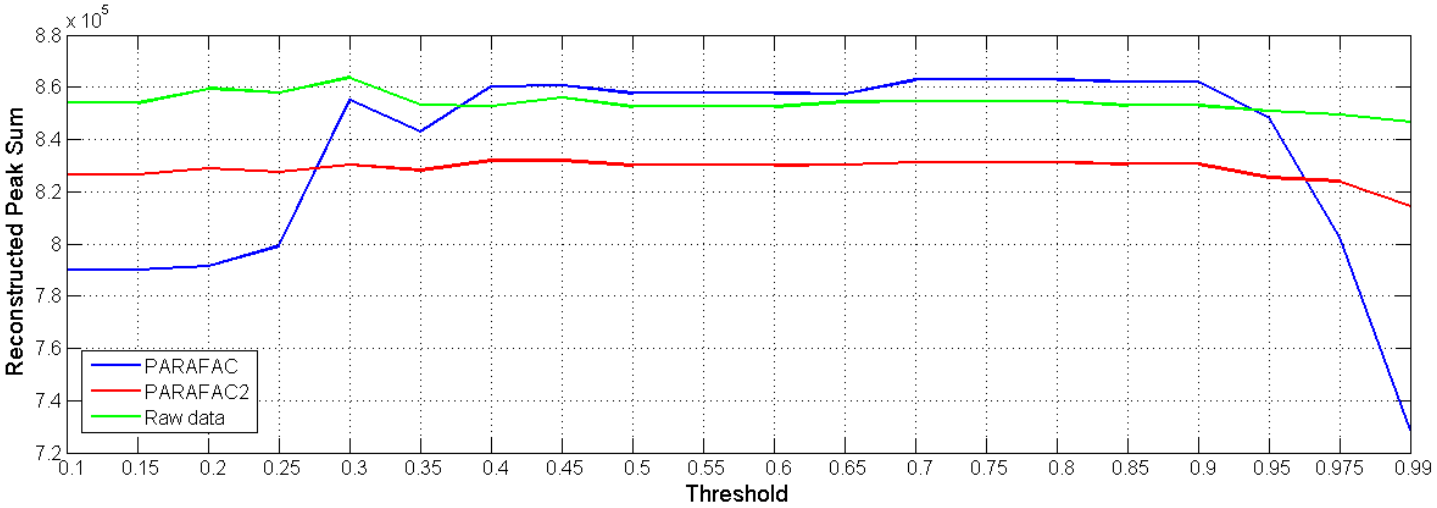


Figure 9. Peak sum from a two-factor **PARAFAC** and **PARAFAC2** model after restoration (Figure 11) of the peak shown in Figure 1. For all peaks the match value was found to be close to or just above 930 indicating that the correct analyte (dimethyl phosphite) has been modeled. The baseline corrected raw data have also been plotted. The relative standard deviation in percent (RSD%) for the peak sum of the raw data and PARAFAC2 were 0.41% and 0.49%, respectively.

Figure 9 shows that, in accordance with the shift corrections in Figure 8, the largest peak sum is achieved when the structure of the data is indeed trilinear (corrected for shift). This is found in the threshold region from 0.40 to 0.9 (Equation 7). So concordance between visual appearance and model derived parameters was achieved.

PARAFAC2 is introduced to show that retention time shift correction is not always a necessity as this shift can be taken care of during the model step and thus, eliminates the need for an additional preprocessing step. In Figure 9 it is expected to see that the peak sum of all peaks from Figure 8 is quite similar as the same analyte is being modeled, although with different degrees of shift. This was confirmed with only small deviations in peak sum for different threshold values (RSD% = 0.49).

Comparing the peak sums of the two multiway models and the raw data indicates that they perform more equally as long as the peaks (first column slices) are aligned. The match values are just above 930 for all peaks restored using a threshold in the interval 0.45 to 0.9 for both PARAFAC and PARAFAC2. As only one sample has been used for this example of threshold values, the consistently slightly higher peak sum for PARAFAC should not be considered significant. Other samples investigated showed a reversed picture or an even smaller difference between the two multiway methods. As a single analyte region should only contain analyte signal and a constant background signal, the use of baseline corrected data for peak integration could be an alternative. This is confirmed in Figure 9. But for the peak sum of the baseline corrected raw data to be valid, it is a prerequisite that other analytes or interferences are not present in the region of interest. Narrowing the region handled a part of this problem for the sample shown in Figure 9, but for other samples interfering peaks were unavoidable increasing the uncertainty of this integration method. Also severe user interaction or complex methods are needed to ensure single analyte regions and thus, this approach is less favorable for finding the peak sum of the analyte of interest in multi-component systems. More importantly, the properties of PARAFAC and PARAFAC2 allow the extraction of the individual contribution of a single analyte if the

unique elution time and mass spectral loadings can be found (proper number of components determined and the correct analyte identified from the match value). Thus, these methods are far more attractive for finding the correct quantitative peak sum values.

4.3 Quantification of bromobenzene

Up until this point, we have demonstrated that peak shifting is a problem if a strict trilinear model should be used and that PARAFAC2 can solve this problem at least at higher signal-to-noise ratios. In this section the effect of shift correction methods followed by PARAFAC will be evaluated and compared to PARAFAC2 over a large range of concentrations of bromobenzene. From the multiway model the peak sum is calculated from Equation 5 and 6 for PARAFAC and PARAFAC2 models, respectively. In general, only models providing a match value above 750 will be considered valid and used for the quantification of bromobenzene. However, models providing a match value below but close to 750 were also considered by plotting model characteristics (not included here) to examine the validity of these models. If valid, the found peak sum will be included in the quantification results.

The capabilities of the models to quantify bromobenzene over a large concentration range were done plotting the logarithm of the peak sum divided by the known concentration versus the logarithm of the known concentration. Assuming that the response to the analyte (i.e., peak sum defined in Equations 5 and 6) is linear with concentration, the peak sum divided by the concentration (i.e., sum/concentration) should yield a constant value for any corresponding sum and concentration pair. In this way slight deviations from linearity at lower concentrations are more easily seen. Quantification of bromobenzene can be seen in Figure 10 for all the individual data arrangements (preprocessing, artificial shifts, different multiway models) presented in Figure 6. Here, the “restored” data refers to the retention time “shifted” data that has been aligned prior to applying PARAFAC.

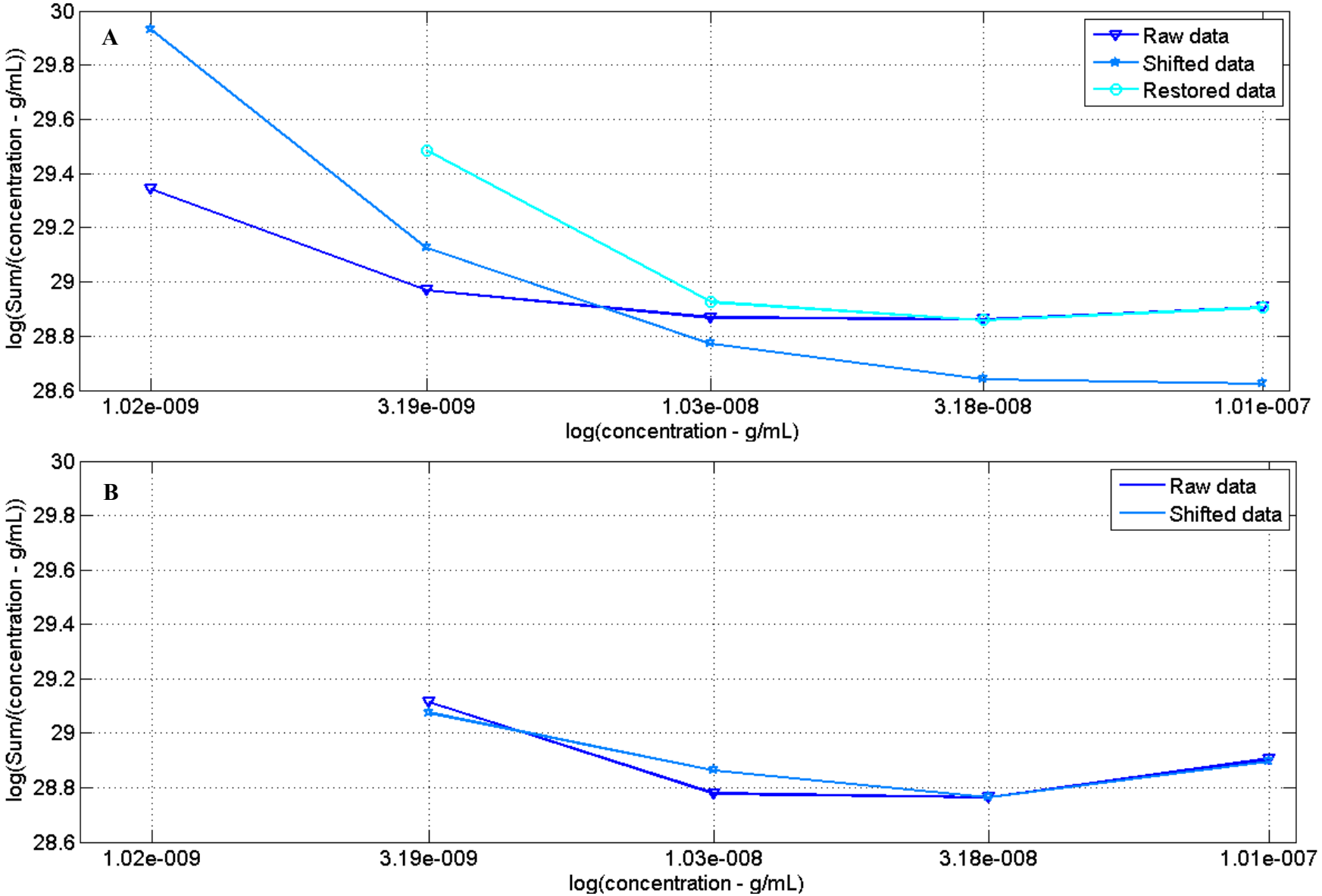


Figure 10. A: **PARAFAC** and B: **PARAFAC2**. Average of replicates of log(Sum/concentration – g/mL) divided by log(concentration – g/mL) for the seven concentrations of bromobenzene. The two lowest concentrations are not detectable in any PARAFAC or PARAFAC2 model and thus, these concentrations are not included. The same axes have been used for both illustrations – the ticks of the abscise axis have been changed to the real concentrations for easier interpretation. Table 3 shows the corresponding RSD% (calculated for peak sum values) for all concentrations and models for the five methods.

Table 3. RSD% values for average peak sum values shown in Figure 11.

		Concentration (g/mL)				
Data	**Model**	**1.01E-7**	**3.18E-8**	**1.03E-8**	**3.19E-9**	**1.02E-9**
Raw data	PARAFAC	2.7	2.6	5.6	5.2	21.8
Shifted data	PARAFAC	2.2	4.7	8.9	42.0	58.3
Restored data	PARAFAC	2.3	3.7	5.8	14.3	
Raw data	PARAFAC2	3.2	4.0	7.8	26.2	
Shifted data	PARAFAC2	3.6	3.2	10.7	44.8	

Figure 10 and Table 3 shows several important findings about the two multiway models categorized in the following. Firstly, the RSD% values are increasing when lowering the concentration (this was expected due to the increased inclusion of noise in the factor modeling). Similar RSD% values have been reported for similar concentrations of bromobenzene [33]. Secondly, for the two largest concentrations all five methods gave similar RSD% values, whereas PARAFAC on the raw data was found to be the most reproducible model. When artificial shifts are applied the peak sum estimation deviates from linearity and the RSD% increases. For PARAFAC2, an increased RSD% was observed for the artificially shifted data at lower concentrations. This suggests that PARAFAC2 not only has more difficulties at lower concentrations, but also that the difficulties depends on the amount of shifts!

4.3.1 Identifying the correct analyte – signal-to-noise ratio

It was expected that the PARAFAC2 modeling of low concentration samples would be a more difficult task compared to PARAFAC. This was confirmed as no PARAFAC2 model was capable of providing a match value above 750 for any replicate in a concentration of 1.02×10^{-9} or below. For two of the shown data arrangements (raw data and shifted data) it was possible using PARAFAC to find the correct analyte (match value above 750) for the 1.02×10^{-9} sample as well.

4.3.2 Shift correction method

The shift correction method presented here was very powerful as shown earlier in Figure 8. However, at low concentrations of bromobenzene the correction power started to decrease as seen from the slight deviation at 3.19×10^{-9} g/mL. This was to some degree expected because the correlation between adjacent slices starts to drop when the signal-to-noise ratio decreases. The mask introduced in the shift correction method ensured that slices were not aligned if there was insufficient evidence of shift present over the mass channels. This situation could result in slices not being aligned even though a shift is observed, but more importantly it reduces the risk of aligning slices having a large perchance correlation coefficient due to artifacts, interfering analytes or noise.

4.3.3 Shifted data

The largest deviation from linearity over the concentration range is observed for PARAFAC modeling of artificially shifted data that has not been aligned. As presented in Figure 7 and 9 PARAFAC has severe problem when shift is observed and this will cause a wrong estimation of the correct peak profile and from this a wrong estimation of the peak area. On the other hand, PARAFAC2 models the shifted data almost as well as it models the raw data. It can be seen that the average over the four replicates for

PARAFAC2 is not affected by the degree of shift as the linearity is similar for both raw and shifted data. However, as stated in Table 3 the RSD% is higher when a larger degree of shift is observed. PARAFAC2 was also applied to the shift corrected data and the results were very similar (results not shown for brevity).

4.3.4 Raw data

Comparing PARAFAC and PARAFAC2 it can be seen that PARAFAC performs better on the raw data, with a slightly better linearity over the four largest concentrations and with the capability of modeling the right analyte at lower concentrations. From Figure 5 is can be seen that a very small shift is present in the bromobenzene peak, but this is found to be so small that PARAFAC is not affected. Several initial runs also indicated that PARAFAC, despite being strictly trilinear, is able to model small systematic shifts in a very efficient way.

4.3.5 What to do for experimentally shifted GC×GC-TOFMS data

The findings from this study suggest that PARAFAC is the most robust model as long as the shift is insignificant between adjacent slices. However, the exact extent of the degree of shift that can be handled adequately by PARAFAC was not further tested here. For shifted data, PARAFAC2 is the obvious choice, but this model was more dependent on the signal-to-noise ratio than was the case for PARAFAC. From these findings the proper data analytical solution (shift correction and/or modeling) for a specific problem at hand can be visualized as follows (Figure 11).

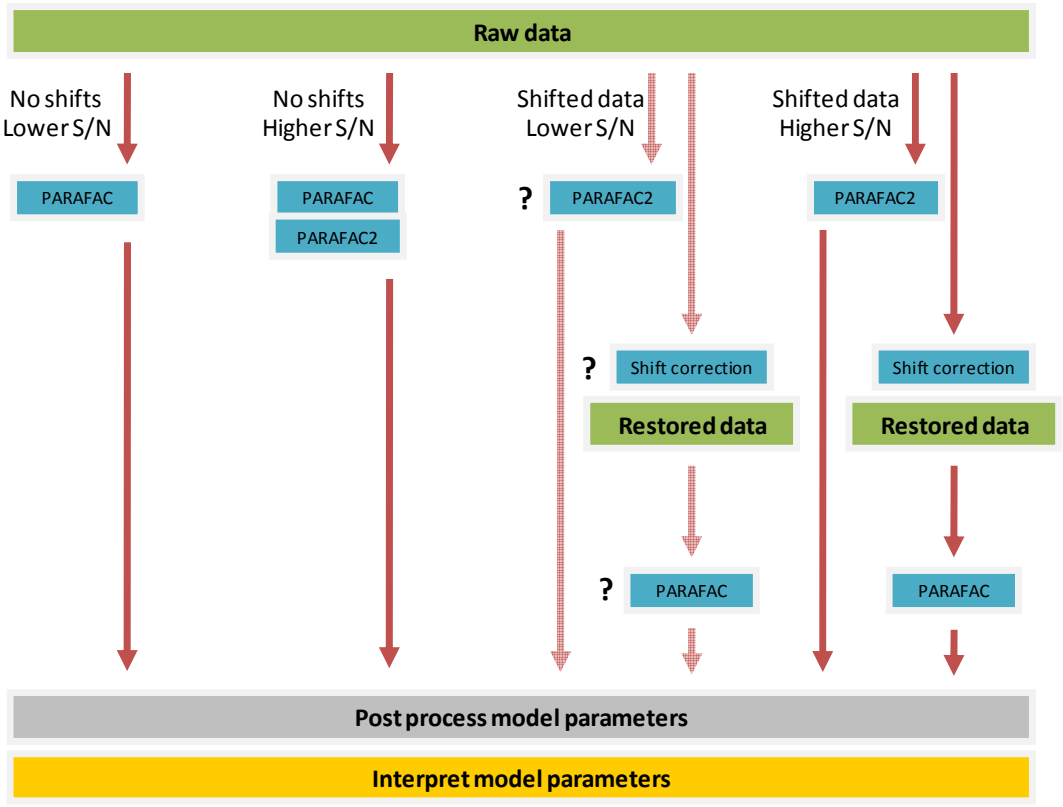


Figure 11. The preferred data analytical way from raw GC×GC-TOFMS data to final model parameters dependent on shifts and signal-to-noise ratios (S/N). The question mark (?) indicates that some considerations are needed prior to the pre-processing and/or modeling as described in the text.

As stated earlier, the analytical solution to a given problem depends mainly on the degree of retention time shift and the signal-to-noise ratio. In the following, the dimethyl phosphite peaks presented in Figure 2 will be evaluated using the two following approaches: (1) Shift correction and PARAFAC, and (2) PARAFAC2. To see the effect of shift correction and modeling over different temperature programs, the samples to be used have the same concentration of dimethyl phosphite and are measured using four different temperature programs (isothermal, 2.5, 5 and 10 ºC/min) in triplicate.

4.4 Evaluation of Dimethyl Phosphite peaks

The dimethyl phosphite peaks run with a temperature program of 2.5, 5 and 10 ºC/min, respectively, (Figure 2) were analyzed in order to show that a realistic severely shifted peak of suitable (here suitable refers to data that is possible to align and where PARAFAC2 is capable of identifying the analyte) signal-to-noise ratio can be modeled equally well using PARAFAC2. The peak restoration for each temperature program used is shown in Figure 12 for the same part shown in Figure 2.

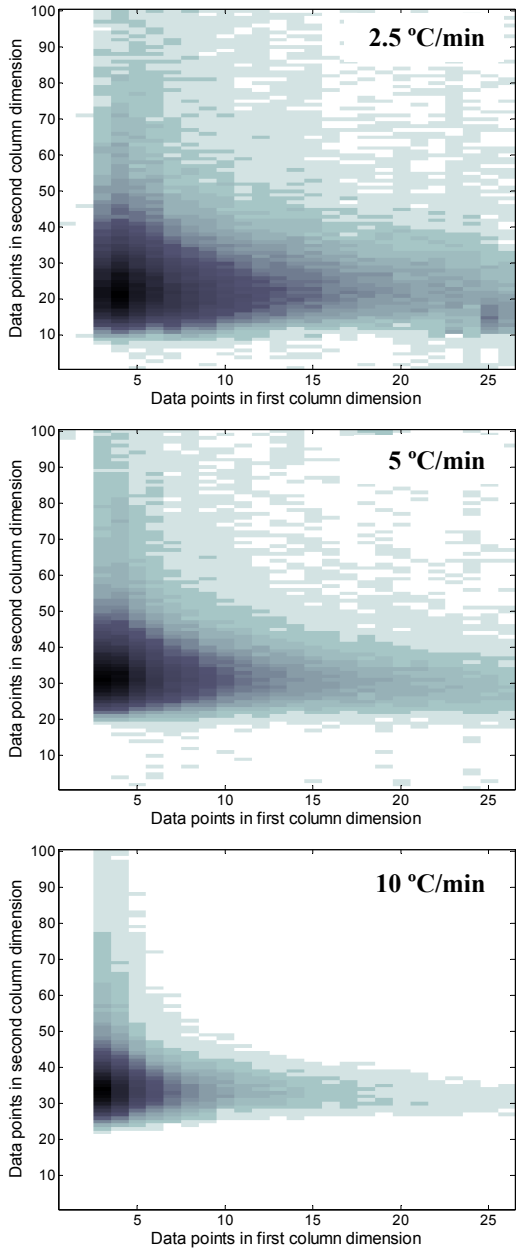


Figure 12. Illustration of shift correction using the method presented in this study. The raw uncorrected data and further information about samples can be seen and found in Figure 2 and the corresponding legend. Notice that the region used for peak sum calculating is larger than shown in the figure. Here only a limited region is shown for better visual confirmation of the restored peak.

Visually, the shift has been corrected, thus indicating that PARAFAC can be applied for the peak modelling as well. To test this, the two challenging solutions were compared in the following by modeling each peak for each temperature program (Tables 4 and 5).

Table 4. **PARAFAC** – data shift corrected and modeled. Comparison of peak sum for samples containing 50 ppm of Dimethyl Phosphite but using different temperature programs. Mean of all samples = 9.0E+5 and RSD = 5.9%.

	Temperature program, ºC/min			
Replicate #	**Isothermal**	**2.5**	**5**	**10**
1	-*	-*	9.0E+05	1.0E+06
2	8.8E+05	8.7E+05	9.0E+05	9.8E+05
3	8.6E+05	8.3E+05	9.3E+05	8.8E+05
Mean	**8.7E+05**	**8.5E+05**	**9.1E+05**	**9.6E+05**

*One replicate was missing.

Table 5. **PARAFAC2** – data modeled. Comparison of peak sum for samples containing 50 ppm of Dimethyl Phosphite but using different temperature programs. Mean of all samples = 9.0E+5 and RSD = 8.2%.

	Temperature program, ºC/min			
Replicate #	**Isothermal**	**2.5**	**5**	**10**
1	-*	-*	8.7E+05	1.0E+06
2	9.5E+05	8.4E+05	8.7E+05	1.0E+06
3	8.6E+05	7.9E+05	9.0E+05	9.1E+05
Mean	**9.0E+05**	**8.1E+05**	**8.8E+05**	**9.9E+05**

*One replicate was missing.

The peak sums from the two models and the different temperature programs shown in Table 4 and 5 should be similar, as all samples have been analyzed at the same concentration of ~50 ppm. As seen, the mean values for both methods are very similar (both within and between models) with a corresponding rather low RSD%. This indicates that no matter which temperature program is used and which method is applied afterwards for the pre-processing and modeling, the results were similar. This supports the findings from the bromobenzene study; the found peak sum is indeed valid no matter the amount of shifting in the data when using shift correction plus PARAFAC or PARAFAC2 alone. This assumes that the signal-to-noise ratio is above a certain threshold and from this study a concentration above 1E-8 g/mL was found to provide accurate and reliable results. No further investigation of the lower limit of signal-to-noise ratio (or concentration) for the dimethyl phosphite data was conducted, but both multiway models were capable of finding the specific analyte in all samples. This was also expected as the concentration of dimethyl phosphite was 5E-5 g/mL and well above the approximate limit of 1E-8 g/mL found in the bromobenzene study where deviations from linearity were observed.

5. Conclusions

The quantitative analysis (or peak signal determination) of GC×GC-TOFMS peaks using trilinear multiway methods, such as PARAFAC, has been documented to be a very efficient technique. Under ideal chromatographic conditions no deviations from trilinearity are expected for single analyte systems. Indeed, our experience with this instrument from prior work is that deviations from trilinearity, while they are observed and should be dealt with as outlined herein, they do not comprise a major fraction of the analyte peaks in a given sample run. However, using temperature programming to cope with the general elution problem, deviations from trilinearity are more likely to be seen for highly retained compounds on the first column, or with fast temperature program rates. Such deviations, seen as shifted peak profiles have been characterized, illustrated and explained by key terms from the chromatographic theory.

Several approaches to correct for and handle deviations in second column elution time for successive first column fractions have been put forward. Depending on the degree of shift and mainly on the signal-to-noise ratio two approaches were superior: (1) Shift correction followed by PARAFAC, and (2) PARAFAC2. They provided a good agreement between peak area (peak sum of elements in the reconstructed analyte peak) from an estimated model and the known concentration. Also the potential of the data analysis approaches were illustrated for artificially shifted peaks at a wide concentration span and for real severe shifts observed at different temperature programs. PARAFAC was found to be more robust at lower signal-to-noise ratios and was capable of detecting and modeling the target analyte at lower concentration than PARAFAC2. However, PARAFAC2 proved to be a very good alternative by removing the requirement of an alignment step before modeling.

In this study only one analyte (i.e., one peak) was of interest at a time by focusing on a small part of the total GC×GC landscape. For this, the presented integer shift correction method showed good potential in correcting shifted profiles. This means that the method with shift correction followed by PARAFAC method cannot directly be transferred to a case with two or more analytes without some consideration. Firstly, more analytes can have different degrees of shift along the second column and thus, might need a more flexible shift correction method. This can be solved by introducing a segment wise interpolation based shift correction method such as Correlation Optimized Warping [55]. PARAFAC2, not suffering from an essential initial shift correction, would be an optimal solution for a multipeak system with different degree of shifts. Secondly, the used match value for model validation must be extended to work with mass spectra from two or more analytes. However, the use of other validation methods such as core consistency and residuals could be included instead.

References

[1] Liu, Z., and Phillips, J. B. (1991) J. Chromatogr. Sci. 29, 227-231
[2] Gorecki, T., Harynuk, J., and Panic, O. (2004) J. Sep. Sci. 27, 359-379
[3] Phillips, J. B., and Beens , J. (1999) J. Chromatogr. A 856, 331-347
[4] van Deursen, M., Beens , J., Reijenga, J., Lipman, P., Cramers, C., and Blomberg, J. (2000) J. High. Resolut. Chromatogr. 23, 507-510
[5] Fraga, C. G., Prazen, B. J., and Synovec, R. E. (2000) Anal. Chem. 72, 4154-4162
[6] Prazen, B. J., Johnson, K. J., Weber, A., and Synovec, R. E. (2001) Anal. Chem. 73, 5677-5682
[7] Mondello, L., Casilli, A., Tranchida, P. Q., Dugo, G., and Dugo, P. (2005) J. Chromatogr. A. 1067, 235-243
[8] Wang, M., Marriott, P. J., Chan, W., Lee, A. W. M., and Huie, C. W. (2006) J. Chromatogr. A. 1112, 361-368
[9] Mayadunne, R., Nguyen, T., and Marriott, P. J. (2005) Anal. Bioanal. Chem. 382, 836-847
[10] Venkatramani, C. J., Xu, J., and Phillips, J. B. (1996) Anal. Chem. 68, 1486-1492
[11] Bruckner, C. A., Prazen, B. J., and Synovec, R. E. (1998) Anal. Chem. 70, 2796-2804
[12] Beens , J., Blomberg, J., and Schoenmakers, P. J. (2000) J. High Res. Chromatogr. 23, 182-188
[13] Dalluge, J., van Rijn, M., Beens , J., Vreuls, R. J. J., and Brinkman, U. A. Th. (2002) J. Chromatogr. 965, 207-217
[14] Mondello, L., Casilli, A., Tranchida, P. Q., Presti, M. L., Dugo, P., and Dugo, G. (2007) Anal. Bioanal. Chem. 389, 1755-1763
[15] Sinha, A. E., Johnson, K. J., Prazen, B. J., Lucas, S. B., Fraga, C. G., and Synovec, R. E. (2003) J. Chromatogr. A 983, 195-204
[16] Johnson, K. J.; Prazen, B. J.; Olund, R. K.; Synovec, R. E. (2002) J. Sep. Sci. 25, 297-303.
[17] van Mispelaar, V. G.; Tas, A. C.; Smilde, A. K.; Schoenmakers, P. J.; van Asten, A. C. (2003) J. Chromatogr. A 1019, 15-29.
[18] Kong, H.; Ye, F.; Lu, X.; Guo, L.; Tian, J.; Xu, G. (2005) J. Chromatogr. A 2005, 1086, 160-164.
[19] Xie, L.; Marriot, P. J.; Adams, M. (2003) Anal. Chim. Acta 500, 211-222.
[20] Peters, S.; Vivó-Truyols, G.; Marriot, P. J.; Schoenmakers, P. J. (2007) J. Chromatogr. A 1156, 14-24.
[21] Sinha, A. E.; Fraga, C. G.; Prazen, B. J.; Synovec, R. E. (2004) J. Chromatogr. A 1027, 269-277.
[22] Sinha, A. E.; Hope, J. L.; Prazen, B. J.; Nilsson, E. J.; Jack, R. M.; Synovec, R. E. (2004) J. Chromatogr. A 1058, 209-215.
[23] Dallüge, J.; van Rijn, M.; Beens, J.; Vreuls, R. J. J.; Brinkman, U. A. Th. (2002) J. Chromatogr. A 965, 207-217.
[24] Lu, X.; Cai, J.; Kong, H.; Wu, M.; Hua, R.; Zhao, M.; Liu, J.; Xu, G. (2003) Analytical Chemistry 2003, 75, 4441-4451.
[25] Focant, J.; Sjodin, A.; Turner, W. E.; Patterson, D. G., Jr. (2004) Anal. Chem. 76, 6313-6320.
[26] Song, S.; Marriott, P.; Kotsos, A.; Drummer, O. H.; Wynne, P. (2004) Forensic Sci. Int. 143, 87-101.
[27] Hope, J. L.; Sinha, A. E.; Prazen, B. J.; Synovec, R. E (2005) J. Chromatogr. A 1086, 185-192.
[28] Jover, E.; Adachour, M.; Bayona, J. M.; Vreuls, R. J. J.; Brinkman, U. A. Th. (2005) J. Chromatogr. A 1086, 2-11.
[29] Moeder, M.; Martin, C.; Schlosser, D.; Harynuk, J.; Górecki, T. (2006) J. Chromatogr. A 1107, 233-239.
[30] Mohler, R. E.; Dombek, K. M.; Hoggard, J. C.; Young, E. T.; Synovec, R. E. (2006) Anal. Chem. 78, 2700-2709.

[31] Pierce, K. M.; Hoggard, J. C.; Hope, J. L.; Rainey, P. M.; Hoofnagle, A. N.; Jack, R. M.; Wright, B. W.; Synovec, R. E. (2006) Anal. Chem. 78, 5068-5075.
[32] Sinha, A. E., Hope, J. L., Prazen, B. J., Fraga, C. G., Nilsson, E. J., & Synovec, R. E. (2004) J. Chromatogr. A, 1056, 145-154
[33] Hoggard, J. C. and Synovec, R. E. (2007) Anal. Chem. 79: 1611-1619
[34] Sinha, A. E., Prazen, B. J. and Synovec, R. E. (2004) Anal. Bioanal. Chem.378 (8): 1948-1951
[35] Fraga, C. G., Bruckner, C. A. and Synovec, R. E. (2001) Anal. Chem. 73 (3): 675–683
[36] Bylund, D., Danielsson, R., Malmquist, G. and Markides, K. E. (2002) J. Chromatogr. A 961 (2): 237-244
[37] Bro, R. (1997) Chemom. Intell. Lab. Syst. Systems 38 (2):149-171
[38] Harshman, R. A. & Lundy, M. E. (1994). Comput. Stat. Data Anal. 18: 39-72
[39] Kiers, H. A. L., Ten Berge, J. M. F. and Bro, R. (1999) J. Chemom. 13 (3-4): 275-294
[40] Bro, R., Andersson, C. A. and H.A.L. Kiers (1999) J. Chemom. 13 (3-4): 295-309
[41] Skov T. and Bro. R. (2005) Sens. Actuators, B 106 (2): 719-729
[42] Wise, B. M., Gallagher, N. B. and Martin. E. B. (2001) J. Chemom. 15 (4): 285-298
[43] Garcia, I., Ortiz, M. C., Sarabia, L. and Aldama, J. M. (2007) Anal. Chim. Acta 587 (2): 222-234
[44] Ovejero-Lopez, I. Bro, R. and Bredie, W. L. P. (2005) Food Qual. Pref. 16 (4): 327-343
[45] Amigo, J. M., Skov, T., Coello, J., Maspoch S. and Bro, R. Submitted for publication in TrAC, Trends Anal. Chem.
[46] Pierce, K. M.; Wood, L. F.; Wright, B. W.; Synovec, R. E. (2005) Anal. Chem. 77: 7735-7743.
[47] Zhang, D.; Huang, X.; Regnier, F. E.; Zhang, M. (2008) Anal. Chem. 80, 2664-2671.
[48] Bro, R. (2006) Crit. Rev. Anal. Chem. 36 (3-4): 279-293
[49] Skov, T. and Bro, R. (2008) Anal. Bioanal. Chem. 390 (1):281-285
[50] Harshman R.A. and Lundy, M.E. (1984) The PARAFAC model for three-way factor analysis and multidimensional scaling, in "Research methods for Multimode data analysis". (Eds. H.G. Law, C.W. Snyder, J.A. Hattie and R.P. McDonald) Praeger, New York.
[51] Bro, R. and Kiers, H. A. L. (2003) J. Chemom. 17 (5): 274-286
[52] Stein, S.E. (1999) J. Am. Soc. Mass. Spectrom., 10 (8): 770-781
[53] Watson, N. E., Siegler, W. C., Hoggard, J. C. and Synovec, R. E. (2007) Anal. Chem. 79 (21): 8270-8280
[54] Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Windig W. and Koch, R. S. (2006) PLS Toolbox 4.0, Eigenvector Research Inc., Manson, WA, 2006.
[55] Tomasi, G., van den Berg, F. and Andersson, C. (2004) J. Chemom. 18 (5): 231-241