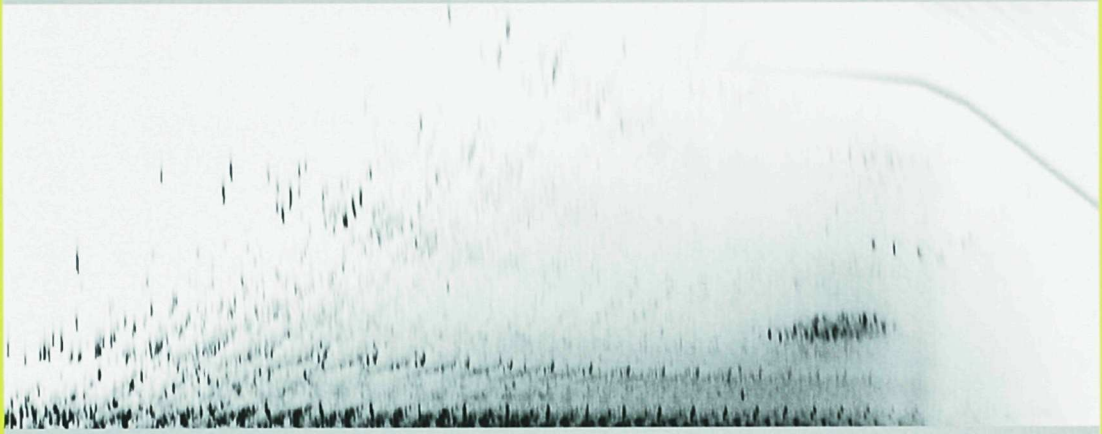
Chromametrics



Valentijn van Mispelaar

CHROMAMETRICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus,
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 15 juni 2005, te 12.00 uur

door VALENTIJN GERARDUS VAN MISPELAAR

geboren te Amersfoort

Promotor:	prof. dr. ir. P.J. Schoenmakers
Co-promotor:	prof. dr. A.K. Smilde
Overige leden:	prof. dr. ir. H.-G. Janssen
	prof. dr. Th. Hankemeier
	dr. A.C. Tas
	dr. J. Blomberg
	dr. W.Th. Kok

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Publication of this thesis was financially supported by TNO Quality of life, Zeist.

Layout:	Valentijn van Mispelaar
Typesetting:	LaTeX
Design Cover:	Josien Geerdink
Printing:	Universal Press, Veenendaal

Contents

List of Figures

List of Tables

Chapter 1

Introduction to chromametrics -Combining chromatography and chemometrics.

1.1 Chromatography and MVA

Even though chromatography is often considered to be a mature technique, this does not mean that new developments do not emerge. By far the most important development in gas chromatography in the last decade has been the introduction of comprehensive two-dimensional gas chromatography by John Phillips [1–3]. This technique separates all sample components according to two independent, or orthogonal, separation mechanisms [4]. Two different GC columns are used in GC×GC. The first-dimension column is (usually) contains a non-polar stationary phase, separating components largely based on their vapour pressures (boiling points). The second-dimension column is considerably smaller (smaller diameter, shorter length) than the first-dimension column, so that separations in the second dimension are much faster. The stationary phase is selected such that this column separates on properties other than volatility, such as molecular shape or polarity. The two columns are coupled using a so-called modulator. This device continuously traps and releases small portions of the effluent. With each modulation, a new second-dimension chromatogram is started. The detector, which is positioned at the end of the second-dimension column, records these fast chro-

matograms. The detector output at the end of a chromatographic run is a large string of second-dimension chromatograms. After "demodulation" [5], a three dimensional chromatogram (two retention axes and an intensity axis) results. The term "chroma^{2}gram " can be used for what is usually represented by a colour or contour plot. A large variety of applications has been described in literature and several review articles discussing GC×GC [3,6–9] have been published.

Comprehensive two-dimensional gas chromatography or GC×GC offers many advantages in comparison with conventional one-dimensional gas chromatography. The main advantages are summarized below:

-GC×GC provides a much larger peak capacity. This can be used globally, for separating very complex examples, or locally for separating analytes from each other or from matrix components. In this context GC×GC can be advantageous, as soon as more than just a few peaks need to be separated.

-GC×GC provides structured separations, if the separation dimensions (separation mechanisms) match the sample dimensions (most relevant structural features of the sample).

-GC×GC provides an increased sensitivity for quantitative analysis.

The first advantage can be used to achieve a better separation between the target components ("analytes") and the surrounding matrix, *i.e.* to increase the analytical selectivity. The second advantage implies substantial benefits for the separation of component groups. The third advantage is facilitated by the modulator, which enables an increase in sensitivity of a factor 4-5.The benefits of GC×GC (or for that matter, any new development) must, therefore, be categorized into each different application of gas chromatography. Fortunately, the large number of individual applications can be reduced to only three generic application types. By doing so, the benefits of GC×GC (and other technological advances in chromatography) can be discussed for each of the three application types. Practical users of chromatography can use this classification scheme to assess the benefits of any new development for their specific application. Comprehensive two-dimensional separation gas chromatography is capable of very impressive separations. However, the resulting chromatograms have a corresponding complexity and size. Dedicated strategies to retrieve information from these highly complex chromatograms have to be considered. Multivariate-analysis techniques may offer such an approach. The use of multivariate-analysis techniques (sometimes referred

to as chemometrics) on chromatographic data is not new.
Multivariate-analysis (MVA) techniques have provided tools for data pre-processing, classification, calibration, and for many other purposes. Well known examples are principal-component analysis (PCA) [10] and partial-least-squares regression (PLS) [11]. The former is often applied to complex data, with the aim of reducing the number of relevant variables, while the latter generally is used to relate measured data to product properties. These techniques facilitate the processing of complex data. Many years ago it was already recognized that the combination of chromatographic separations and MVA techniques offered excellent possibilities for the characterization of (complex) samples. Already in the mid-sixties, the first references on the combination of MVA and chromatography appeared. However, not all of these references can easily be retrieved. The biannual reviews in Analytical Chemistry provide a useful historical overview of the combination of the two fields. Due to the large number of references in the literature, ranging from well-respected journals to rather obscure sources, it is difficult (if not impossible) to give a comprehensive overview of all the work performed in this field. However, the individual references can be divided into a limited number of common research topics or categories of applications.
First of all, MVA techniques have been applied to the detection side of the separation system. Examples of such applications are the deconvolution of mass-spectrometric (MS) [12–14] or photo-diode array (PDA) [15] data obtained after chromatographic separations. MVA techniques are used for calculating the 'pure' component profiles, thus mathematically separating components that were not completely resolved by chromatography. This approach makes use of the so-called "second-order advantage" [15], which implies that a complete spectrum, rather than a single data point, is obtained at any one time. Another example in this category concerns the enhancement of signal-to-noise ratios [16]. Very early examples date back as far as 1974 [17].
The second category of applications of MVA techniques in chromatography concerns (quantitative) structure - retention relationships (QSRR). In the large field of quantitative structure - activity relationships (QSAR), relationships are sought between molecular structure and (biological) activity. Examples of QSRR include the relationship between the structure of anti-malarial drugs and their LC retention factors [18], and the identification of

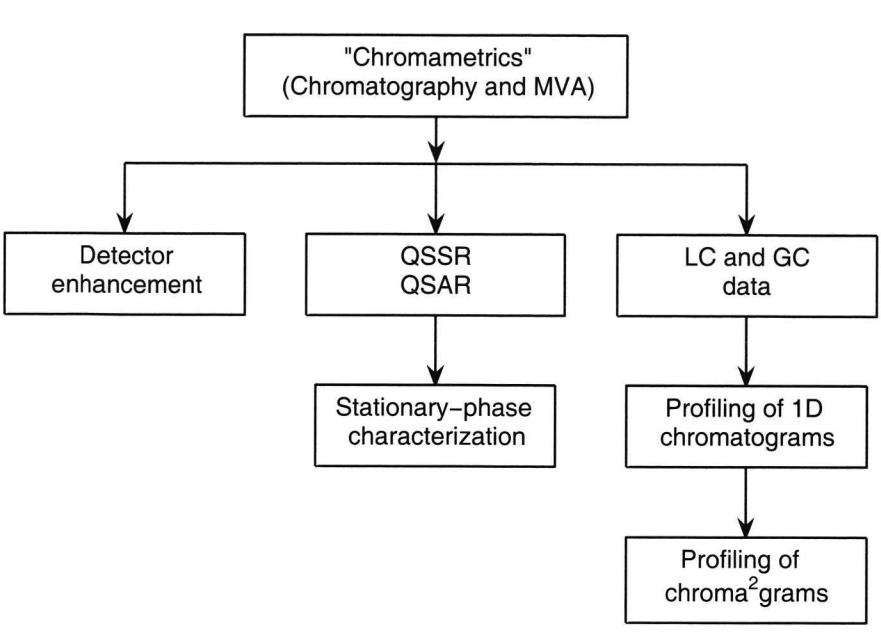


Figure 1.1: Breakdown of MVA applications into various fields.

structural features related to the retention mechanism in HPLC [19, 20] or GC [21].

The third category of applications is the multivariate comparison of chromatographic profiles. Either chromatography-derived information (*e.g.* peak tables) or chromatographic profiles can be used for this purpose. Applications in both liquid and gas chromatography can be found in the literature. Early examples of the classification of chromatographic data date back several decades [22] and there is now a large variety of applications, such as the classification of brain tissue [23], PCB analysis [24–26], fatty acids [14, 27], petroleum-based accelerants [28], fuel-spills [29], jet fuels [30], wine [31], coffee [32], and pheromones [33, 34]. The prediction of product properties using MVA tools and gas chromatographic analysis has also been described for various types of applications, such as fuel performance [35] and octane numbers [36].

One of the largest limitations for the application of MVA techniques to chro-

matographic data is the occurrence of retention-time shifts. Chromatographic methods will always feature some variation along the time axis, due to instrumental (variations in flow and temperature) and fundamental (adsorption isotherm) reasons. Multivariate-analysis methods will consider retention-time changes as changes in chemical composition. Elimination or at least reduction of these shifts is, therefore, of prime importance. Understandably, much attention has been focused on this problem [37–40]. Whereas conventional, high-resolution, one-dimensional gas chromatography allows several hundreds of (equally spread) peaks to be baseline separated, GC×GC has a peak capacity which is an order of magnitude higher. This is obviously very favourable from a chromatographic point-of-view, but advanced data-analysis tools become mandatory for handling such complex data. For this reason, several articles have already appeared that describe the application of MVA in combination with two-dimensional separation techniques. The

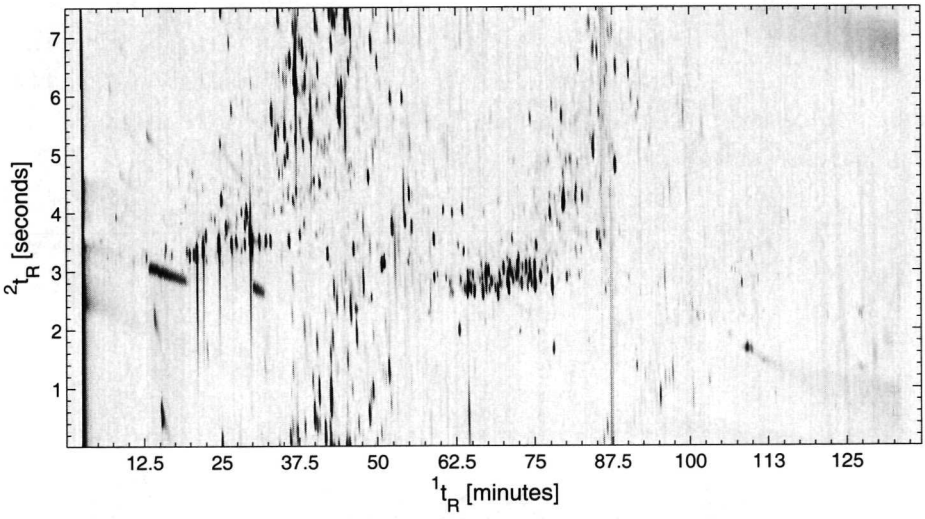


Figure 1.2: Chroma^{2}gram of lavender oil.

second-order advantage of GC×GC has been exploited by Bruckner *et al.* [13] for accurately determining the concentrations of (partially) overlapping components. An additional benefit of the MVA approach was the enhanced signal-to-noise ratio in comparison with other quantification methods. The same group used multiway models (parallel factor analysis or "Parafac") for the de-

convolution of data obtained using GC×GC in combination with time-of-flight mass spectrometry (GC×GC-TOF-MS) [41]. GC×GC already provides second-order data. The third-order advantage of a mass spectrometric detector is used here for improved mass-spectral selectivity. The potential of obtaining highly detailed fingerprints by GC×GC is illustrated by the separation of essential oils from lavender, bergamot and ylang-ylang. The chroma^{2}gram of a lavender-type essential oil is presented in Figure 1.2. Similar chromatograms were recorded for two other types of lavender oils, as well as for two types of bergamot oil and one type of ylang-ylang oil. All samples were analyzed in triplicate. In addition, a 1/1 (v/v) mixture of Bergamot O and Lavender S was prepared and analyzed. By considering each chromatogram as a fingerprint of the essential oil, comparisons between the products can be made. In this case the 'inner-product correlation' [42], a matrix equivalent to the correlation coefficient, was used to calculate similarities between the chromatograms. For a set of 20 chroma^{2}grams, a correlation matrix of 400 correlation values can be constructed (each sample correlated to all 20 samples). Since such data are difficult to interpret, the data matrix was forced into a two-dimensional representation using multidimensional scaling [43]. Results are shown in Figure 1.3, where the essential oils are clustered based on their chemical fingerprints.

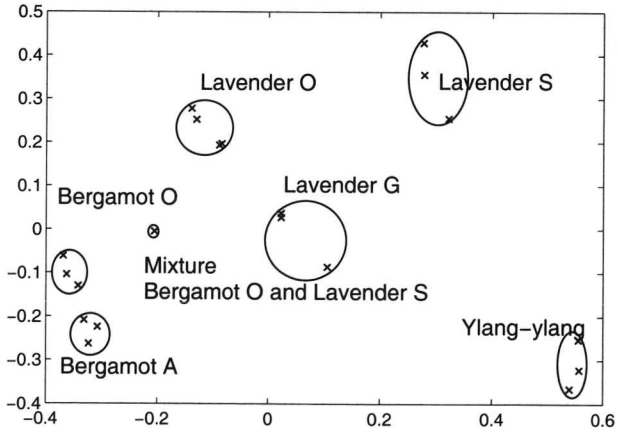


Figure 1.3: Multidimensional scaling of correlation matrix calculated for 20 essential-oil samples.

1.2 Aims and scope of this thesis

The aims of this thesis are to explore and extend the possibilities of multivariate-analysis techniques applied to comprehensive two-dimensional gas chromatographic separations. In Chapter 2, a classification scheme is presented by which three generic types of applications of gas chromatography (GC) and comprehensive two-dimensional gas chromatography (GC×GC) can be distinguished. These generic application types allow virtually any (gas) chromatographic application to be classified. This aids scientists on the forefront of technology to judge the merits of technological advances for different applications. For the practical users of chromatography, this scheme helps to judge the usefulness of new developments for their particular application. The Chapters 3-6 in this thesis are arranged according to this classification scheme.

Chapter 3 deals with so-called target-compound analysis. Multiway methods are used for fast quantification of a limited number of predefined components.

Chapter 4 describes tools for the quantification of component groups. Tools such as baseline correction and splining are described.

Chapter 5 describes the use of an alignment strategy for two-dimensional separations. This alignment technique applies image-processing tools for identifying identical points (or landmarks) in two different images (chroma^{2}grams in this case). The selected points form the basis for a second-order polynomial function describing the difference between the two images.

Chapter 6 describes the classification of crude oils using GC×GC and MVA techniques. An objective variable-selection technique is used to discriminate between "informative" and "non-informative" data.

The techniques described in these Chapters can be considered to be generic 'chromametric' tools, which facilitate the extraction and interpretation of information from highly complex chroma^{2}grams.

Chapter 2

Classifying Chromatographic Applications.*

For practical chromatographers it is extremely difficult to judge the merits and limitations of new technological developments. On the other hand it is nearly impossible for those at the forefront of technology to judge the implications of their efforts for all specific applications of chromatography. Both chromatographers and researchers can be aided by a classification of the numerous specific applications into a few well-defined categories. In this Chapter, we propose such a classification of all chemical analyses by chromatography into three generic types of applications, *viz.* target-compound analysis, group-type separation, and fingerprinting. The requirements for each type are discussed in general terms. The classification scheme is applied to assess the benefits and limitations of comprehensive two-dimensional gas chromatography (GC×GC) and the possible additional benefits of using multivariate-analysis (MVA) techniques for each type of application. The conclusions pertaining to the generic types of applications are indicative for the implications of new developments for specific chemical analyses by chromatography.

*Published as: *A Novel System for Classifying Chromatographic Applications, Exemplified by* GC×GC *and Multivariate Analysis*, V.G. van Mispelaar, H.G. Janssen, A.C. Tas and P.J. Schoenmakers in: Journal of chromatography A **1071** (2005), 229-237.

2.1 Introduction

Chromatography nowadays is widely used, with numerous applications in a wide range of application areas. Liquid and gas chromatography (LC and GC, respectively) are often said to be mature techniques. Indeed, reliable methods and instruments are available and the techniques can be applied by trained analysts, as well as by skilled experts. However, the word "mature" by no means implies that there are no more developments in the area. For example, in LC new column concepts (*e.g.* monolithic columns [44] and chips [45]) are developing strongly and instrumentation is progressing towards higher pressures [46] and two-dimensional analyses [47, 48]. In GC, comprehensive two-dimensional separations form the most striking example. For the practical user of chromatography it is increasingly difficult to judge the merits of new developments for her or his application. New techniques and methods are generally illustrated in the literature by one or a few specific applications. For example, in his pioneering paper on GC×GC, John Phillips showed the benefits of the technique only for petrochemical products [1, 2]. Almost all work in the first six years of GC×GC was restricted to this application area. A commonly voiced misconception during this time was that GC×GC was only applicable to petrochemical products. It was not until 1997, after Phillips had published the separation of PCB's [49], that the technique slowly started to be adopted in other application areas.

Another example is the introduction of narrow-bore GC for fast separations. Initially, the method was used incorrectly, which significantly delayed its acceptance [50]. Although narrow-bore capillary columns are an excellent means for speeding-up GC separations, they are not suitable for all applications. For a while, fast GC in general and narrow-bore columns in particular suffered from a bad reputation. The eventual acceptance of fast GC was aided by a series of review articles [50–52], describing the various options for faster separations and strategies for selecting the optimal approach.

As stated before, it is not always easy for chromatographers in practice to judge the benefits of new developments for their applications. When developing new instruments and techniques, it is also impossible to establish the advantages and limitations for each single application of chromatography. Fortunately, in practice this will hardly be necessary. We believe that by looking at commonalities between applications, the almost infinite num-

ber of applications can be reduced to a small number of generic application types. In this contribution, we will describe a novel scheme for classifying chromatographic applications. All chemical analysis (*viz.* qualitative and quantitative analysis) of chromato-graphy are divided in three categories. For each of these application types, the general merits (and limitations) of new developments can easily be identified. This allows a rapid assessment of the value of new developments for each specific application of chromatography. We do not consider applications other than chemical analysis, such as measurements of physical properties by, for example, size-exclusion chromatography.
Two recent technological advances in chromatography, comprehensive two-dimensional gas chromatography (GC×GC) as such, and GC×GC in combination with multivariate analysis (MVA), will be used to demonstrate the proposed strategy. The advantages of these developments for the various types of applications will be described. Before we can do so, the relevant aspects of these new technologies must be briefly described.

2.1.1 Comprehensive two-dimensional gas chromatography

The concept of GC×GC was pioneered and advocated by John Phillips [1–3]. A typical GC×GC system consists of two chromatographic columns in series. These columns separate components according to two different properties. Between the first- and second-dimension columns, a modulator is located. Small portions of the effluent from the first-dimension column are continuously trapped and released by this device. The result of a comprehensive two-dimensional separation can be visualized as a two-dimensional chromatogram, extending into three dimensions (two retention-time axes and an intensity axis). This technique provides an unsurpassed peak capacity and, as a result, very detailed chromatograms (so-called chroma^{2}grams).

2.1.2 Multivariate analysis

Multivariate-analysis (MVA) techniques are chemometric tools for retrieving information from very large datasets, which are too complex for human interpretation. MVA techniques aim to reduce the data complexity. They result in strongly simplified representations of the data. In general, MVA techniques can be divided into two categories:

Projection techniques for the visualization of differences or similarities between the samples. The best-known example is principal-component analysis (PCA) [10]. Since in many cases objects are described by (many) highly correlated variables, the dimensionality of the dataset is reduced if these variables can be replaced by a small number of principal components. Each sample in the dataset is then described by a number of principal-component loadings (profiles in which the original variables are expressed) and principal-components scores (weight factors for each loading). The resulting projection provides a much clearer picture of the dataset and allows the selection of relevant variables. When differences between classes of samples are desired, discriminant-analysis techniques, such as principal-component-discriminant analysis (PCDA) [53] can be used.

Calibration techniques to establish relationships between measurements and, for example, product behaviour. Regression and calibration techniques aim to correlate the dataset with one or more external variables. For example, in an industrial process the water content of a product can be a very important specification. By continuous monitoring the process using near-infrared spectroscopy (NIR), a set of spectra is collected. By applying a multivariate-calibration technique, the water content in newly measured samples can be predicted, based on a previously constructed calibration model. Well-known examples of these techniques are principal component regression (PCR) and partial least squares (PLS) [11].

2.2 Theory

As stated in the introduction, we believe that all chromatographic applications can be classified into a small number of generic application types. The approach we propose here starts from the way in which the chromatographic signal is converted into the desired information on the sample. In our philosophy, only three translation strategies are applied. This implies that we distinguish only three generic types of applications in chromatography.

2.2.1 Type I: Target-compound analysis

The most-common type of application is based on converting retention times into peak identities and the corresponding peak areas into amounts or con-

centrations. The actual information desired, are the concentrations of a finite number of pre-specified components. This strategy is generally referred to as "target-compound analysis". The important keywords for this generic type of application are described below.

Isolation (local resolution): The compounds of interest ("targets") must be sufficiently separated from each other and from the sample matrix. Separation of other compounds present in the matrix is not required. The apparent resolution of target compounds may be enhanced by using specific detectors.

Identification: Obviously, unambiguous identification is very important in this type of application. Retention times (or Kovats-Indices) are useful in this respect. However, only specific detectors (particularly mass spectrometry) can provide irrefutable proof of compound identity.

Reliable calibration: After recording the chromatographic signal, the peak areas must be transformed into concentrations. This can be achieved by calculating calibration factors from pure standards or reference materials. This requires the compounds to be stable and available in pure form. If this is not the case, FID response factors can be estimated using the theory of Scanlon and Willis [54].

Sensitivity: In order to analyze low levels of compounds, a sensitive chromatographic system is required. This can be achieved by using sensitive detectors, suitable methods of sample preparation, and/or large-volume injection.

2.2.2 Type II: Group-type analysis

In the second type of application, component groups are of interest, rather than individual components. This is, for example, the case when there is a strong correlation between the levels of specific component-classes and the relevant product-properties or if a particular group of components is toxic. Instead of "component groups", the name "pseudo-components" is also used. Pseudo-components often have structural properties in common, such as specific end-groups, an identical number of aromatic rings, a specific configuration of double bonds, *etc.*. Separation of the samples into individual component groups (or separating component groups from the matrix) provides valuable information. This strategy can be referred to as "Group-

type analysis". The main requirements for this type of application are the following.

Group-type selectivity: Separation between the different component groups or between the component group(s) and the matrix is required. Separation within the groups is generally not necessary or even undesirable.

Quantitative detection: Because the goal of this type of application is to obtain quantitative results on groups of components, a quantitative detector is required, which offers an equal response for all members of a component group. Whereas mass spectrometry may be an excellent tool for structure elucidation, it often fails in providing quantitative results on groups of components, due to large differences in ionization-efficiencies between components in a group and other reasons.

Group identification: Unlike in Type-I (target-compound) applications, where only a limited number of individual peaks in the chromatogram are relevant, component groups have to be identified and quantified. Therefore, this type of application requires group-wise integration and quantification methods.

2.2.3 Type III: Fingerprinting

In Type-I and Type-II applications, prior knowledge on the sample is required, *i.e.* the components or component groups of interest are known a priori. This is not always the case. A typical example is a product - that for unknown reasons - does not meet its specifications (in other words, it is "off-spec"). Such products may contain unknown components, which are responsible for the failure. In these situations, there will then be an urge to identify the responsible component(s) or component groups. One approach may be to quantify all components present in the sample and to correlate the results with the product properties. In most cases, this approach will be very demanding, if not impossible. A different approach is to consider the entire chromatogram as a "fingerprint" of the sample. By correlating this fingerprint with the product properties, component(s) or profiles can be traced to the off-spec condition. This approach heavily relies on MVA techniques. The requirements for Type-III (Fingerprinting) applications are the following.

Peak capacity: Since each component present in the sample is potentially

relevant, systems with a very high peak-capacity are required to separate as many components as possible.

Retention-time stability: Since MVA techniques generally require large sets of data and since recording chromatograms requires a considerable amount of time, ensuring system stability is a formidable challenge. Even minor shifts in retention times may render an entire dataset useless.

Detector stability: Analogous to retention-time stability, detector response should be very stable over time. Otherwise, erroneous conclusions may be drawn.

Dynamic range: Since both major and minor components can be relevant, a wide dynamic range is required.

Multivariate-analysis techniques: In order to correlate fingerprints with certain product properties, multivariate-correlation techniques are required. Examples are (PLS) and Principal-Components Regression (PCR).

The result of a "Fingerprinting" application may be a set of peaks or a group of peaks that correlates with a certain product property, it may be a (multivariate) classification of samples, a library of chromatograms, *etc.*. Identification of the identified (pseudo-) components will turn a "Fingerprinting" application into a target-compound (Type I) or group-type (Type II) application. Table 2.1 gives an overview of the three types of applications distinguished, summarizing also the main requirements for each application type.

Application Type I: Target-compound analysis	Application Type II: Group-type analysis	Application Type III: Fingerprinting
Target compounds isolated ("local resolution")	Group selectivity	High peak capacity
Unambiguous identification	Separation between groups	High retention-time stability
Reliable calibration	Quantitative detection	Stable response
High sensitivity	Group identification	Broad dynamic range
	Group quantitation	Multivariate-analysis tools

Table 2.1: Overview of requirements for the three application types.

2.3 Results

Chromatographic separations are performed to obtain information on specific samples. In the theory section, three ways of translating the chromatogram into the desired information have been discussed, which resulted in three types of applications. New developments in chromatography generally result in more or better information from faster, simpler, or cheaper methods. The consequences of such developments for each type of application can be very different. This can, for example, be illustrated by discussing the introduction of a new column with a different selectivity in GC. In case of a target-compound analysis in a very complex sample, the column will probably be of little use. On the old column, certain target components probably co-eluted (mutually or with matrix components), whereas other co-elutions are likely to occur on the new column. Multi-dimensional operation of the old and the new column may result in improved target-compound analysis, but only at the expense of increased efforts and analysis time. For group-type separations (Type II), however, the new column could be very interesting. Below, the application-type concept will be used to discuss the merits of two recent developments in chromatography, *viz.* comprehensive two-dimensional gas chromatography (GC×GC) and its combination with MVA.

2.3.1 Target-compound analysis (Type I)

Since many target-compound analysis focus on very complex materials, there is a perpetual effort to develop separation systems capable of separating target components from one another and from the matrix. In many cases, the resulting chromatographic methods are related to product specifications, process control, environmental issues, legislation, *etc.*. According to the requirements mentioned in the theory section of this paper, new developments that are useful for this type of application should provide adequate local resolution (peak capacity), unambiguous identification, and adequate sensitivity.

With respect to the **isolation of target compounds in the chromatogram**, GC×GC is superior to conventional 1D-GC. This may substantially aid the separation of target components from each other and from surrounding matrix peaks. With respect to **unambiguous identification**, GC×GC offers two retention coordinates instead of one. This im-

proves the accuracy of peak assignment. However, there still is no accepted two-dimensional alternative to the one-dimensional Kovats retention index. Moreover, coupling to MS requires very fast MS instruments (*e.g.* time of flight). Also, GC×GC-TOF-MS yields massive amounts of data. This makes the analysis and interpretation of GC×GC-TOF-MS data much more difficult than in the case of GC-MS. Finally, peak-compression provides an increase in **sensitivity**, typically by a factor of 4 or 5 in comparison with conventional 1D-GC [55]. The application of MVA techniques has already proven advantageous for Type-I applications of GC×GC. Fraga *et al.* have reported the use of the generalized-rank-annihilation method (GRAM) for lowering the detection limits and resolving overlapping peaks [16]. Enhanced productivity may be a second advantage of the application of multivariate-analysis methods. In Chapter 3 the describes the of of so-called multiway methods for the rapid quantification of large datasets is described.
To illustrate the merits of GC×GC for Type-I applications, the analysis of key flavour ingredients in a vanilla extract is used as an example. This application requires a truly high-resolution GC system.

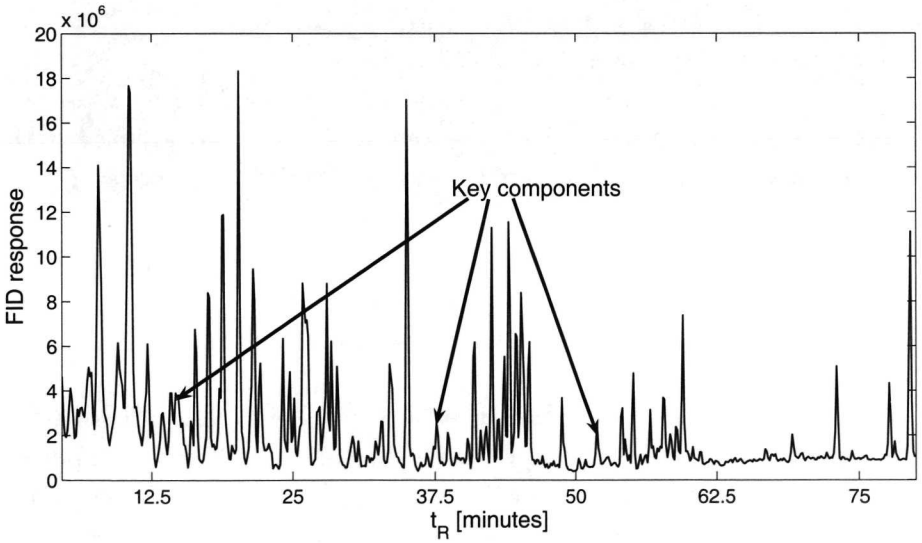


Figure 2.1: Separation of a vanilla extract using 1D-GC.

Figure 2.1 shows the chromatogram of a vanilla sample. The indicated key components appear more-or-less separated from the matrix. A chroma^{2}gram

of the same vanilla sample, however, gives a better impression of the true complexity. The sample is clearly much more complex than suggested by conventional 1D-GC.

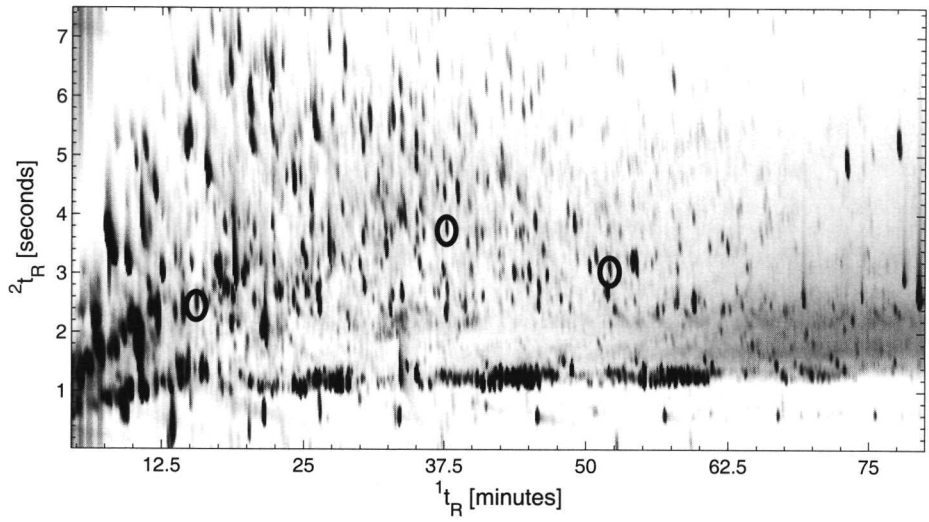


Figure 2.2: Chroma^{2}gram of vanilla extract. Circles indicate components of interest.

Components in Figure 2.2, which are on the same vertical line as the indicated target compounds, would co-elute in the corresponding one-dimensional chromatogram. In this example,conventional 1D-GC would clearly overestimate the concentration of the key vanilla components. It is for this reason that target-compound analysis in general (and within the flavour and perfume fields in particular) are often performed using GC-MS [56]. GC×GC-TOF-MS combines many of the advantages of GC×GC and GC-MS for Type-I applications. Arguably, it is the best separation technique currently available [57]. Other examples in the literature of "target-compound-analysis" by GC×GC include biomarkers in oil [58], key flavour compounds in essential oils [59, 60], doping control [61], garlic-flavour analysis [62], and pesticides in food extracts [63].

GC×GC is extremely useful for Type-I applications. However, it is not always the preferred method. For relatively simple samples (*e.g.* homologous series), the components can be separated in one dimension. For instance,

Fraga *et al.* reported the separation of a seven-compound mixture (alkyl benzenes) using GC×GC [16]. Although they described a nice demonstration of the applicability of chemometric methods for quantification purposes, the separation of such simple mixtures could probably also be achieved on a one-dimensional separation system.

2.3.2 Group-type analysis (Type II)

Many complex chemical and natural materials contain huge numbers of individual components. In general, the latter belong to only a limited number of chemical classes. A group of components belonging to one class is often referred to as a pseudo-component. For pseudo-component analysis, it is common practice in gas chromatography to first separate samples into as many components as possible, followed by grouping of the components belonging to each class. The final results are usually the concentrations of one or more components groups, rather than the concentrations of individual components. Pseudo-components can be related to sample properties, such as hydrogen conversion in hydrocarbon mixtures, toxicity in PCB containing samples, the degree of unsaturation of fatty acids, *etc.*.

The first Type-II applications of GC×GC have been reported in the field of petrochemical analysis [64]. Although these products virtually always contain an overwhelming number of components, the number of chemical classes is much-more limited. Structured separations are obtained by GC×GC, which substantially aids component identification [65]. In terms of the sample-dimensionality theory of Giddings [66], the two separation dimensions are chosen so as to match the most significant sample dimensions (*e.g.* volatility and polarity).

By far the main benefit of GC×GC for Type-II applications is the possibility to obtain **structured chromatograms**. By matching the separation dimensions with the sample dimensions, component groups actually elute in bands parallel to the first dimension axis. In the theory section, three requirements were addressed for Type II applications: selectivity, quantitative detection and group-wise integration. With respect to **selectivity**, GC×GC provides excellent possibilities. Since the first and second dimensions generally involve columns coated with different stationary phases, components are separated according to two different (sets of) properties. An important

possibility is the decoupling of volatility and polarity contributions to analyte retention [65]. Due to peak compression in the modulator, GC×GC has a minor advantage over conventional 1D-GC with respect to **quantitative detection**. The requirement for **group-wise integration** can - in principle - easily be met in GC×GC. The result of an ordered separation may be that components are grouped in classes. Therefore, group-wise integration can be achieved by drawing boxes around component groups. A summation within such a group yields a "group area", as described in Chapter 4. Chemometric methods may help to assign chromatographic peaks to component groups or with the deconvolution of (partly) overlapping component -groups. However, no publications have addressed these possibilities so far. In order to illustrate the advantages of GC×GC for Type-II applications, the group-type analysis of petrochemical products is used as an example. Traditionally, group-type analysis of light hydrocarbon fractions is achieved using multi-dimensional column-switching GC. GC×GC has proven to be a successful alternative.

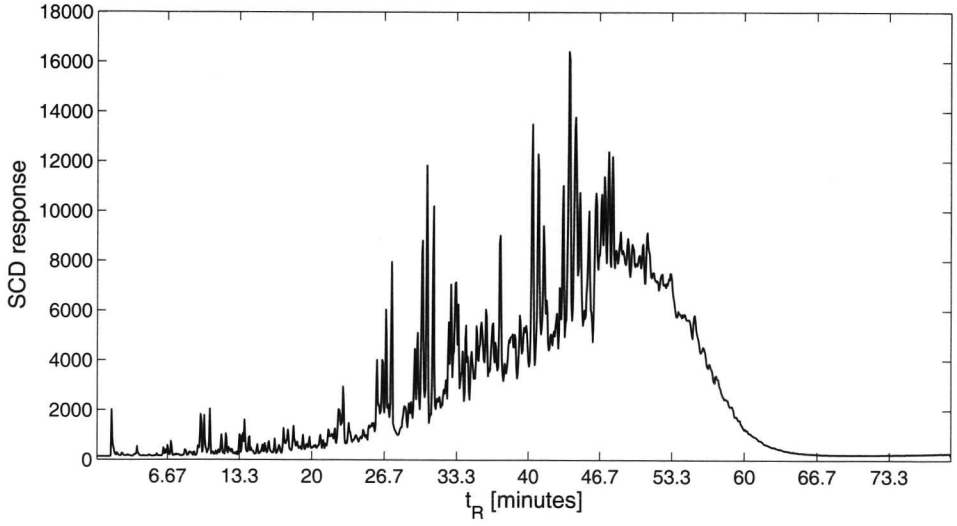


Figure 2.3: One-dimensional chromatogram of cycle-oil obtained with GC-SCD.

Figure 2.3 shows the one-dimensional chromatogram of a cycle-oil obtained with sulphur-chemiluminescence detection (SCD). Although present capillary GC columns have an impressive separation power, they are not really adequate for such complex samples.

The combination of columns coated with different stationary phases in heart-cutting multi-dimensional GC is of rather limited value for group-type separations [67]. The combination of a boiling-point separation in the first dimension and a polarity separation in the second dimension in a GC×GC system results in a highly ordered chromatogram, in which the various pseudo-components can be distinguished. In Figure 2.4, the comprehensive two-dimensional separation of a cycle oil with GC×GC-SCD is shown.

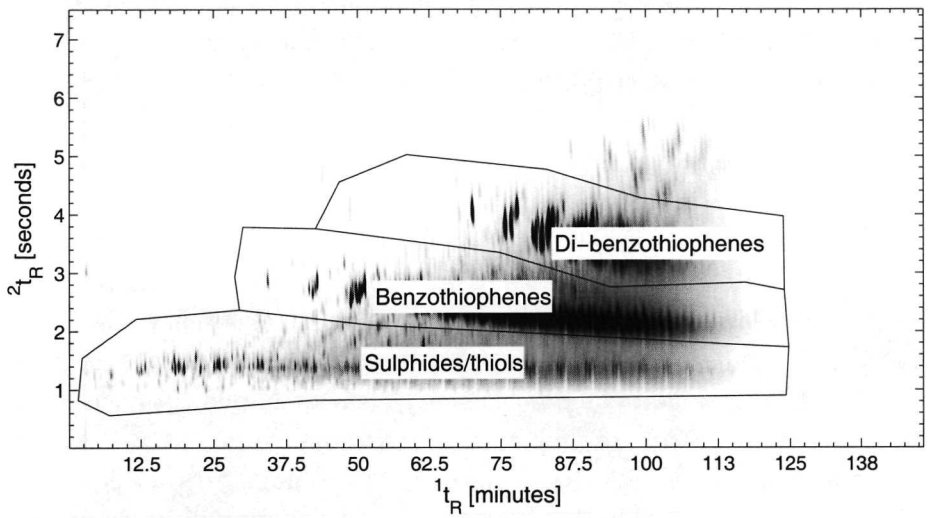


Figure 2.4: Group-type separation of a cycle-oil with GC×GC-SCD.

The boxes indicate the regions in which specific compound groups elute. These regions can also be used for quantitative purposes.
Other applications of Type-II analysis by GC×GC are the determination of the degree of unsaturation of fatty acids [68, 69] and the classification of PCB's according to planarity [70].

2.3.3 Fingerprinting (Type III)

One specific research area that thrives on the "fingerprinting" approach is the identification of "biomarkers" (or "disease markers") in systems biology. In this application area, the correlation between sick and healthy patients and their metabolomic profiles needs to be established. This is achieved by

analyzing samples from sufficiently large numbers of "test subjects" (human, animal, or vegetable) of known condition (either suffering from a particular disease or syndrome, or not). Correlations between the chromatographic profiles and the status of the objects can be established using pattern-recognition tools. This allows the identification of biomarkers for a particular disease, which can then be used to detect diseases at an early stage or to assess the effectiveness of drug treatments. The field of proteomics relies heavily on this approach [71]. In the theory section, the requirements for Type-III applications have been identified as peak capacity, retention-time stability and dynamic-range. With respect to **peak capacity**, GC×GC provides roughly the product of the peak capacities of the first- and second-dimension columns. This is a much higher number than what can be obtained in conventional, one-dimensional chromatography. GC×GC hence clearly facilitates the recording of detailed fingerprints of complex materials. For the second requirement, **retention-time stability**, the problems are aggravated in GC×GC in comparison with conventional 1D-GC. In GC×GC separations, retention-time shifts can occur in both dimensions. This makes data pre-processing a formidable challenge for GC×GC. Fortunately, developments in both GC instruments and column technology have resulted in much-more-stable instruments. With respect to the **dynamic range**, GC×GC suffers from the application of (relatively) narrow-bore columns in the second dimension. Narrow-bore, thin film columns have a low sample capacity and can compromise the wide dynamic range of the applied detectors, such as FID and MS.

The use of MVA techniques is often needed for this type of application. Since even conventional 1D-GC is able to generate hundreds of peaks, conventional interpretation does not allow a fast correlation between sample composition and product properties. In many cases, a combination of components can be correlated with product performance, patient status, *etc.*. Univariate methods are not able to deduce highly correlated component profiles. Multivariate-analysis methods can, however, be used, since they are highly suitable for reducing the complexity of the datasets. In two-dimensional electrophoresis, this approach has, for example, been used to classify maps of lymphomas [72].

For successful multivariate analysis, data-pre-processing techniques (such as scaling, aligning, and variable selection) are obligatory to overcome, for ex-

ample, retention-time shifts. Fingerprinting applications using MVA of conventional 1D-GC have hardly been described. Publications in this field concern the prediction of mineral-oil properties based on gas-chromatographic separations [28], the detection of the origin of fuel spills [73], and metabolic profiling with GC-MS [74]. For the combination of GC×GC with MVA techniques, hardly any references can be found [75, 76].

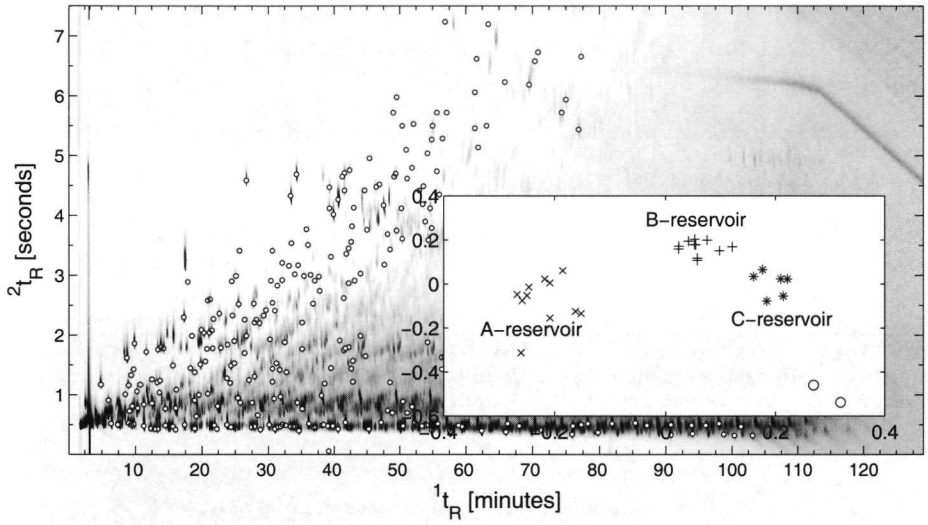


Figure 2.5: Clustering of crude oils according to their origin using GC×GC data.

However, the combination of GC×GC and MVA is potentially very powerful, since the fingerprints obtained from GC×GC contain very much information. To fully exploit this potential, powerful data pre-processing techniques are needed. Below, we will illustrate the power of MVA methods using an example from oil production. Differentiation between highly similar crude-oil reservoirs (*i.e.* wells within one oil field) is very difficult, but vital for monitoring the oil production. GC×GC provides very detailed chromatograms with up to 6000 components. The challenges for chromatography and MVA of such samples and data are formidable. Every chromatogram represents a very large dataset. This means that many samples are typically required to obtain a representative impression. Moreover, the comparison of samples is hindered by retention-time shifts and by imperfections in the integration.

Variable-selection techniques have been used to reduce the dataset to approximately 300 components. Although it is quite feasible to separate 300 peaks in one-dimensional GC, the 300 peaks from GC×GC are pre-selected for relevance and absence of interference from irrelevant peaks. The selected components were subjected to a discriminant analysis, resulting in the clustering of the samples into three reservoirs (A, B and C) (described in Chapter 6 of this thesis). Figure 2.5 shows the GC×GC chromatogram of a crude oil indicating the peaks that are used to build a discrimination model. Table 2.2 summarizes the requirements for each application type and lists examples of published applications.

Application	Type I: Target-compound analysis	Type II: Group-type analysis	Type III: Fingerprinting
Multivariate analysis (MVA)	Component assignment Component alignment Quantification	Group assignment Group alignment Group quantification	Preprocessing (alignment, scaling) Classification and clustering methods
Application examples	PCB's Key flavour components	Cis/trans classification Hydrocarbon-group type analysis	Metabolomics Crude-oil clustering

Table 2.2: MVA requirements and application examples of GC×GC in combination with MVA for the three generic application types

2.4 Discussion and conclusion

All applications of chromatography can be classified into three generic types of applications: target-compound analyses, group-type separation and fingerprinting. The implications of new technological developments can be rigorously assessed at the generic level. The general benefits and limitations for each application type can be translated into practical advantages and disadvantages for the numerous specific applications of chromatography. The classification scheme should aid the developers of new technologies to understand and explain the potential of their products to the chromatographic community. It should also aid practical chromatographers in understanding the implications of new developments for their specific applications. The proposed approach has been used to assess the merits of

	Application requirements	(Dis-)advantages of GC×GC	Additional (dis)advantages of MVA
TYPE I: Target-compound analysis	High peak capacity	Much higher peak capacity	Possible deconvolution of overlapping peaks
	Reliable component identification	Two retention axes	Possible correction for retention time shifts[a]
	Reliable quantification	Greater reliability due to less peak overlap	Possibility of deconvolution
	Adequate Sensitivity	Peak compression	Signal/noise filtering
Type II: Group-type analysis	Selectivity	Structured chromatograms; Decoupling of analyte parameters (e.g. volatility and polarity)	Group-deconvolution
	Constant detector response within group	N/A	N/A
	Group-quantification	Structured separations; Less peak overlap	Potentially very much faster quantitation
TYPE III: Fingerprinting	Peak capacity	Much higher peak capacity	Data-reduction and clustering techniques
	Retention-time stability	Retention shifts may occur in two dimensions	Possible correction for retention time shifts[a]

[a] During pre-processing stage.

Table 2.3: MVA requirements and application examples of GC×GC in combination with MVA for the three generic application types

GC×GC, and the additional advantages of its combination with MVA. For each of the three generic types of applications, clear benefits and limitations could be identified and recommendations for specific applications could be deduced. Table 2.3 reviews the advantages and disadvantages of GC×GC- as a stand-alone application or in combination with MVA techniques - in comparison with conventional 1D-GC.

Acknowledgements

We are very grateful to J. Blomberg of the Shell Research and Technology Centre (Amsterdam, The Netherlands) for providing the data of Figures 2.3 and 2.4 and for many stimulating discussions.

Chapter 3

Quantitative GC×GC analysis.*

Quantitative analysis using comprehensive two-dimensional gas chromatography is still rarely reported. This is largely due to a lack of suitable software. The objective of the present study is to generate quantitative results from a large GC×GC dataset, consisting of thirty-two chromatograms. In this dataset, six target components need to be quantified. We compare the results of conventional integration with those obtained using so-called "multiway analysis methods". With regard to accuracy and precision, integration performs slightly better than Parallel Factor (Parafac) analysis. In terms of speed and possibilities for automation, multiway methods in general are far superior to traditional integration.

3.1 Introduction

The demand for reliable, precise and accurate data in the analysis of complex mixtures is rapidly increasing. This is partly caused by an increased demand for comprehensive characterization of mixtures due to legislation, health concerns, controlled processing, *etc.*. Meeting this demand requires significant technological advances.

*Published as: *Quantitative analysis of Target Compounds by Comprehensive Two-Dimensional Gas Chromatography*, V.G. van Mispelaar, A.C. Tas, A.K. Smilde, A.C van Asten and P.J. Schoenmakers in: Journal of chromatography A **1019** (2003), 15-29.

One of the greatest and most significant advances for the characterization of complex mixtures of volatile compounds is comprehensive two-dimensional gas chromatography (GC×GC). This technique was pioneered and advocated by the late John Phillips [1–3]. In GC×GC, two GC columns are used. The fist-dimension column is (usually) a conventional capillary GC column, with a typical internal diameter of 250 or 320 μm. Most commonly, this column contains a non-polar stationary phase, so that it separates components largely based on their vapour pressures (boiling points). The second-dimension column is considerably smaller (smaller diameter, shorter length) than the first-dimension column, so that separations in the second dimension are much faster. The stationary phase is selected such that this column separates on properties other than volatility, such as molecular shape or polarity. Between the two columns, a modulator is placed. In the modulation process, small portions of the effluent from the first-dimension column are accumulated and injected into the second column. A large number of fractions are collected and the resulting gas chromatogram contains a large series of such fast chromatograms in series (and partly superimposed). When the second-dimension chromatograms are 'demodulated' [5], a two-dimensional representation of the separation is obtained and typically displayed as a colour or contour plot, a so-called chroma^{2}gram.

Many applications have shown the advantages of GC×GC over conventional GC, for instance in the petrochemical field [64, 77], essential oil [59, 60], fatty acids [69], pesticides [78], and polychlorinated biphenyls [50]. However, GC×GC is still largely a method for qualitative analysis. Quantitative analysis by GC×GC is much less commonly used. The first quantitative results obtained with GC×GC were reported by Beens *et al.* [79] in 1998. They applied an in-house integration package called "Tweedee" for the characterization of heavy gas oils. This program integrated 2D slices, followed by a summation along the first dimension. The program worked well on baseline-separated peaks, but it lacked sophisticated integration algorithms to cope with less-ideal situations. Several research groups working on GC×GC have developed their own software for quantification [80, 81].

Synovec *et al.* reported on the use of multiway methods using the so-called "second-order advantage" in order to retrieve quantitative data from GC×GC [15, 16, 76, 82, 83]. Multiway routines, such as the Generalized Rank-Annihilation Method (GRAM) were demonstrated to perform well in this

respect. For the flavour and fragrance industry, quantification of trace com-

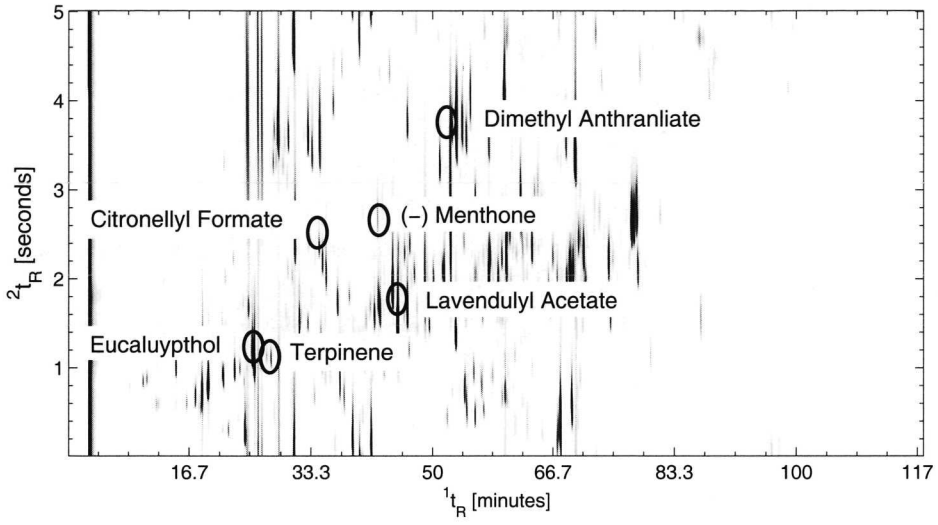


Figure 3.1: Chroma^{2}gram of a (synthetic) perfume sample.

pounds, such as essential-oil markers, is of high importance. The presence of essential oils has a big impact on both the olfactory quality and the price of a perfume. For quality control or competitor analysis, identification and quantification of essential oils is usually done through markers [56]. Cheap and chemically produced alternative ingredients often co-exist in the perfume composition. Markers are present at low levels in the essential oils and thus at trace levels in the entire formulation. GC×GC should yield accurate concentrations and low detection limits for these components.

This study describes the use of GC×GC to quantify essential-oil markers in full perfumes (*i.e.* complete formulations). Our goal has been to quantitate a limited number of target analytes in very complex GC×GC chromatograms by comparing integration with multiway-analysis methods.

3.2 Theory

3.2.1 Quantification

Integration of one-dimensional chromatograms to obtain quantitative data is well established. Typically, first-order and second-order derivatives are used to mathematically detect the peak "start", peak top, and peak "stop", as well as the presence of shoulders. Although far from trivial, integration is now generally regarded as reliable, reasonably fast, and accurate. However, for data obtained from a comprehensive two-dimensional separation, chromatographic integration yields only data that are integrated in the direction of the second-dimension chromatograms. A second step has to be performed to integrate the data along the direction of the first dimension. This can be done either automatically [84] or manually by drawing summation boxes, as is done in the present study.
Another approach can be to utilize the "second-order advantage", using the two-way nature of the measuring techniques. This can be achieved through so-called "multiway techniques", as described below. Synovec and Fraga described the application of the Generalized Rank-Annihilation Method (GRAM) to GC×GC data in order to retrieve both pure-component elution profiles and quantitative information [16, 85].

Nomenclature
In this article, standardized terminology is used, as proposed by Kiers [86] for multiway analysis and by Schoenmakers, Marriott and Beens [87] for comprehensive two-dimensional chromatography.

3.2.2 Multivariate analysis

Standard multivariate data analysis requires data to be arranged in a two-way structure, such as a table or a matrix. An example is a table in spectroscopy, where for different samples absorbances are measured at different wavelengths. The table can be indexed by sample-number and by wavelength and therefore is a two-way array. Two-way methods, such as principal-components analysis (PCA) can be used for the analysis of this type of data. When the relation between absorbances and, for instance, concentrations is wanted, techniques such as Partial Least Squares (PLS)

regression can be used. In many applications PCA and PLS are of prime importance. Near-infrared spectroscopy (NIR) essentially relies on these techniques [88].

In many other cases, a two-way arrangement of the data is not sufficient and a description in more directions is needed. One example is formed by the excitation/emission fluorescence spectra of a set of samples. Each data element can then be indexed by the sample number, emission wavelength, and excitation wavelength, which implies that we have a three-way matrix. When data can be arranged in matrices of order three or higher, it is referred to as "multiway" data. Multiway methods have been applied to a wide variety of problems [89]. Some examples are the decomposition of fluorescence-spectroscopy data of poly-aromatic hydrocarbons [90], the prediction of amino-acid concentrations in sugar with fluorescence spectroscopy [91], data exploration of food analysis with gas chromatography and sensory data [92], and the calibration of liquid-chromatographic systems [93, 94]. A dataset obtained from comprehensive two-dimensional gas chromatography (GC×GC) with flame-ionization detection can also be regarded as three-way. When all second-dimension chromatograms are stacked on top of each other, each data element can be indexed by first, - and second-dimension retention axes and by sample number and contains an FID response. When mass-spectrometry is used, data can be regarded as a four-way arrangement and indexed by first- and second-dimension retention axes, a mass axis and a sample number. Each element then contains an ion count.

Methods for multiway analysis are extensions of existing MVA routines. PCA can be generalized to higher order data in two different ways, Parallel Factor Analysis (Parafac) and Tucker models, while PLS can be expanded, for example, to multilinear PLS [95] or to multiway covariates regression [96].

Parafac

Parallel Factor (Parafac) analysis is a generalization of PCA toward higher orders. It is a true multiway technique, which decomposes a multiway dataset into one or more combinations of vectors ("triads"). The Parafac model was proposed in the 1970's, independently by Carrol and Chang under the name CANDECOMP (Canonical Decomposition) [97] and by Harshman under the name Parafac [98]. Essentially, Parafac models the

data as follows: In this schematic overview, the stacked chromatograms

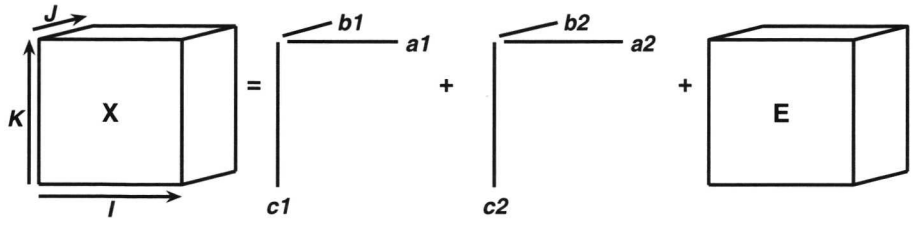


Figure 3.2: Schematic two factor Parafac model.

are represented by the matrix $\boldsymbol{X}$ with dimensions $(I \times J \times K)$. In our case I indicates the first-dimension fraction (retention time), J the second-dimension retention time, and K the specific sample or injection.
Tri-linear decomposition through Parafac into a two-component model yields two triads, $a1$, $b1$, $c1$ and $a2$, $b2$, $c2$ with the dimensions $a(I \times 1)$, $b(J \times 1)$ and $c(K \times 1)$. Matrix $\boldsymbol{E}$ contains the data not fitted in this two-component model. Each coordinate in the data cube $\boldsymbol{X}$ can be described by Parafac as the product of the first- and second-dimension points in both a and b, multiplied by the relative concentration in c:

$$x_{ijk} = \sum_{i=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{3.1}$$

Where:

x_{ijk}	FID response at ${}^1t_{R,i}$ and ${}^2t_{R,j}$ for the k^{th} sample
R	Number of factors (components)
a_{ir}	Value of ${}^1t_{R,i}$ (first-dimension elution time i) for component r
b_{jr}	Value for ${}^2t_{R,j}$ (second-dimension elution time j)for component r
c_{kr}	Relative concentration for sample k and component r
e_{ijk}	Residual for coordinate e_{ijk}

Described in a different (slab-wise) way the Parafac decomposition is given by:

$$\boldsymbol{X}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{B}^T + \boldsymbol{E}_k \tag{3.2}$$

Where:

$\boldsymbol{X}_k$	chromatogram for k^{th} sample $(I \times J)$
$\boldsymbol{A}$	Matrix containing ${}^1t_R$ elution profile $(I \times R)$
$\boldsymbol{D}$	Diagonal containing weights (relative concentrations) of k^{th} sample of $\boldsymbol{X}$ $(R \times R)$ (From $\boldsymbol{C}$)
$\boldsymbol{B}$	Matrix containing ${}^2t_R$ elution profiles $(R \times J)$
$\boldsymbol{E}_k$	Residual for k^{th} sample in $\boldsymbol{X}$ $(I \times J)$

Constraints

In mathematical terms, empirical models are used to describe the data as well as possible. Negative values in the estimated loadings arise if these result in a better solution. However, negative values are often undesirable in chemical and physical applications. In our case, negative FID responses and concentrations are clearly unrealistic. By limiting the solution in the concentration direction to non-negative values, and peak profiles in both retention directions to be unimodal and non-negative, chemically meaningful results are obtained.

Uniqueness

For many bilinear methods there is a problem concerning rotational freedom. The loadings in spectral bilinear decomposition represent linear combinations of the rotated, pure spectra. Additional information is required to find the true (physical) pure-component spectra. Parafac, however, is capable of finding the true underlying pure-component spectra if the dataset is truly trilinear.

The Parafac and Parafac2 equations are solved through an alternating least-squares minimization of the residual matrix and yields direct estimates of the concentrations without bias.

Parafac2

Most multiway methods assume parallel proportional profiles (*e.g.* invariable absorbtion wavelengths or elution times). In some cases, such as batch-process analysis, the time required to process a batch may vary, resulting in unequal record lengths. In chromatography, peaks may shift due to minor deviations in conditions. Many multiway methods cannot deal with such shifted (time) axes. Parafac2 handles shifted profiles through the inner-product structure [99]. It uses this property to deal with stretched

time axes. The Parafac2 algorithm can be described schematically as follows:

$$\boldsymbol{X}_k = \boldsymbol{A}_k \boldsymbol{D}_k \boldsymbol{B}^T + \boldsymbol{E}_k \tag{3.3}$$

Where:

- $\boldsymbol{A}_k$ Matrix containing ${}^1t_R$ elution profile the for k^{th} sample$(I \times R)$
- $\boldsymbol{D}_k$ Diagonal containing weights (relative concentrations) of k^{th} sample of $\boldsymbol{X}$ $(R \times R)$
- $\boldsymbol{B}$ Matrix containing ${}^2t_R$ elution profiles $(R \times J)$.
- $\boldsymbol{E}_k$ Residual for k^{th} sample in $\boldsymbol{X}$ $(I \times J)$.

A useful property of $\boldsymbol{A}_k$ is that $\boldsymbol{A}_k^T \boldsymbol{A}_k = \boldsymbol{A}^T \boldsymbol{A}$ for $k = 1, .., K$. In other words, the cross-product of the $\boldsymbol{A}$ matrix is constant for all samples. In Table 3.1, a simulated GC×GC peak is given $(\boldsymbol{A})$, while $(\boldsymbol{B})$ and $(\boldsymbol{C})$ are the same distribution shifted by one and two positions, respectively. Figure 3.3 projects the data in the form of a two-dimensional peak The inner products ($\boldsymbol{A}^T\boldsymbol{A}$, $\boldsymbol{B}^T\boldsymbol{B}$ and $\boldsymbol{C}^T\boldsymbol{C}$) yield the square of each cell and on the diagonal the sum of squares appears. Note the three situations yield identical values.

In literature, Parafac2 has been used for the decomposition of LC-PDA

$\boldsymbol{A}$				$\boldsymbol{B}$				$\boldsymbol{C}$			
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	1	0	0	0	0	0	0	0	0	0
0	0	2	0	0	0	1	0	0	0	0	0
0	1	3	1	0	0	2	0	0	0	1	0
0	3	5	3	0	1	3	1	0	2	0	0
0	1	3	1	0	3	5	3	0	1	3	1
0	0	2	0	0	1	3	1	0	3	5	3
0	0	1	0	0	0	2	0	0	1	3	1
0	0	0	0	0	0	1	0	0	0	2	0
0	0	0	0	0	0	0	0	0	0	1	0
$\boldsymbol{A}^T\boldsymbol{A}$				$\boldsymbol{B}^T\boldsymbol{B}$				$\boldsymbol{C}^T\boldsymbol{C}$			
0	0	0	0	0	0	0	0	0	0	0	0
0	11	21	11	0	11	21	11	0	11	21	11
0	21	53	21	0	21	53	21	0	21	53	21
0	11	21	11	0	11	21	11	0	11	21	11

Table 3.1: Simulated GC×GC data for Parafac2.

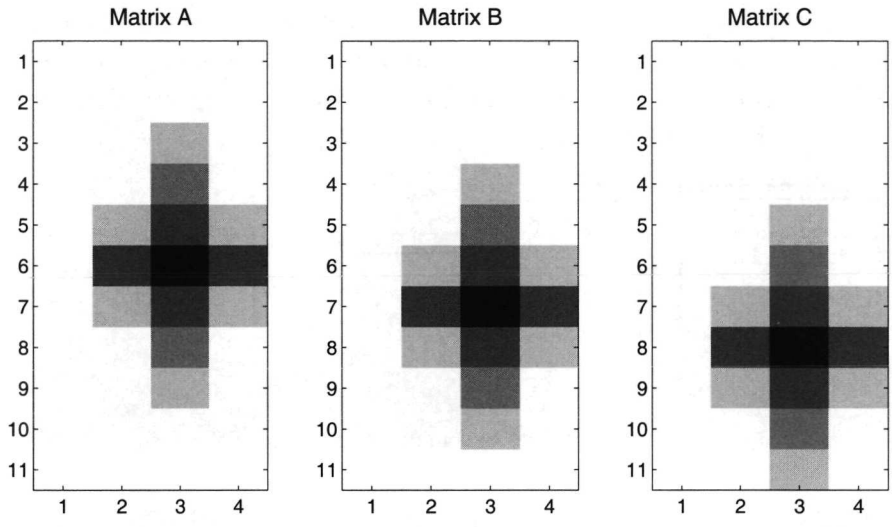


Figure 3.3: Effect of shift of peak position on inner-product.

(Liquid Chromatography - Photo-Diode Array) data [100] and for fault detection in batch-process monitoring [101].
Parafac2 only permits the inner-structure relationship in one direction. For LC-PDA this limitation is easy to justify, as retention-time shifts only occur in the LC direction. For GC×GC, however, shifts can (and will) occur in both retention directions, but they are not identical along the two retention axes. In the second dimension, a peak typically spans at least 15 points, while in the first dimension a maximum of 7 slices encompass a peak. Therefore, the flexibility of Parafac2 is applied along the first-dimension axis, to deal with differences in peak profiles between different injections.

Multilinear PLS

Partial-Least-Squares (PLS) regression is a method for building regression models between independent ($\boldsymbol{X}$) and dependent (y) variables. First, a regression model is calculated, based on calibration data. Decomposition is accomplished in such a way that the computed score vectors of $\boldsymbol{X}$ have maximum covariance with y. Applying the model to samples (unknowns) yields prediction of y.
One specific extension of PLS toward higher orders is called multi-linear

Partial-Least-Squares (NPLS) regression. In this method a multidimensional model is constructed to describe the variance in y. A schematic overview of NPLS is shown below: The NPLS method does not feature built-in con-

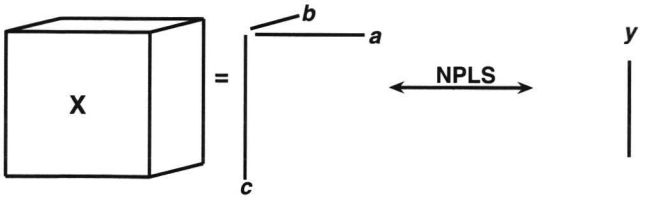


Figure 3.4: Schematic NPLS model.

straints, which may lead to erroneous predictions. Furthermore, in our case the NPLS model needs to be trained using a calibration dataset containing only standards. This may lead to the introduction of additional errors, since the samples contain many more components than the calibration mixtures. Bro has used the NPLS method for the determination of fly ash content in sugar by fluorescence spectroscopy [95] and for the quantification of isomers from tandem-MS experiments [102]. According to the nomenclature of Bro [95], the data presented in the present article can be described by a tri-PLS-1 model (three orders in $\boldsymbol{X}$ and one order in y).
The advantage of NPLS models is their ease of use. The construction of a model is straightforward and there is no external regression step involved. Application of the NPLS method directly yields concentrations for the samples.

3.3 Experimental

3.3.1 Instrumentation

The GC×GC system consists of an HP6890 series GC (Agilent Technologies, Wilmington, DE, USA), configured with a flame-ionization detector (FID) and a Gerstel Cis-4 PTV injector (Gerstel, Muhlheim an der Rühr, Germany) and retrofitted with a second-generation modulator (Zoex, Lincoln, NE, USA) as described by Phillips *et al.* [103]. This device contains a rotating "Sweeper" thermal modulator and a cassette system, which enables

independent heating of the second-dimension column.
The column-set consisted of a 10 m length × 0.25 mm i.d. × 0.25 mm film thickness DB-1 column (J&W Scientific, Folsom, CA, USA). The second-dimension column was 1.2 m × 0.1 mm × 0.1 mm DB-Wax (J&W). The modulation capillary was a 0.07 m × 0.1 mm × 3.5 mm SE-54 column (Quadrex, New Haven, CT, USA). Between the first-dimension column and the modulator, the modulator and the second-dimension column and the second-dimension column and the detector, diphenyltetramethyl-disilazane (DPTMDS) deactivated fused-silica tubing was used (0.1 m × 0.1 mm, TSP 100200-D10, BGB Analytik, Anwil, Switzerland). Columns were coupled with custom-made press-fits (Techrom, Purmerend, The Netherlands).
The carrier gas was helium set at a pressure of 200 kPa, resulting in a flow of approximately 0.8 ml/min at a temperature of 40°C, except for the second calibration mixture, which was analyzed at a carrier gas pressure of 175 kPa, with the intention of inducing retention-time shifts and variations in the first-dimension peak shapes.
The temperature of the first-dimension column oven was programmed from 35°C (5 min isothermal) to 225°C (5 min isothermal) at 2°C/min. The second-dimension column temperature was maintained at 30°C above that of the first-dimension column during the entire experiment.
The modulator was operated at 0.25 rev/s and a slit voltage of 70 V was used (resulting in approximately 100°C elevation of the slotted heater relative to the oven temperature). The modulation time (*i.e.* the time between successive modulations) was 5 seconds.

Instrument control and data processing
The detector signal was recorded with EZ-Chrom Elite software (version 2.61, SP1 SSI, Willemstad, The Netherlands) with an acquisition rate of 50.08 Hz in order to obtain a sufficient number of points across a peak. Data handling was performed with software written in MATLAB R13 (The Mathworks, Natick, MA, USA) running on a Compaq Evo 6000 equipped with two Xeon 2.2 GHz processors and 1 GB RAM. Data-handling routines were developed in-house. In addition, the NetCDF toolbox [104] and the N-way toolbox [105] version 2.10 of the KVL Food-Technology (Department of Dairy and Food Science, Copenhagen, Denmark) were used.

Samples

A set of seven different perfume mixtures for different purposes (detergents and personal care) was selected by Unilever's Perfume Competence Centre (PCC). The samples contained twelve target compounds, but this study is limited to the quantification of essential-oil markers which are γ-terpinene, citronellyl formate, dimethyl anthranliate, lavendulyl acetate, eucalypthol and (-) menthone. The other six components are not reported here for reasons of confidentiality.

The samples were diluted tenfold with 1-propanol (Lichrosolv grade; Merck, Darmstadt, Germany) containing accurately weighted concentrations of approximately 0.25% *n*-decane (Baker grade, min. 99%; Baker, Deventer, The Netherlands) as internal standard. Solutions were prepared in triplicate.

Calibration mixtures of all 12 components were prepared in the same internal-standard solution with concentrations at five levels ranging from 10 to 1500 mg/kg. All calibration solutions were measured in duplicate. To assess the accuracy of the quantification methods, a second calibration mixture was made, containing the same standards, but at concentrations of approximately 200 mg/kg. The calibration mixtures were measured in between the samples. The second calibration standard was measured using a slightly lower carrier gas pressure (175 kPa), forcing retention variations in both the first and second dimensions.

In Figure 3.1 a chroma^{2}gram of a typical synthetic perfume sample is shown. The broad peaks eluting around ${}^1t_R = 25$ min and ${}^2t_R = 3$ to 5 s result from dipropylene-glycol, which is used as an odourless solvent in the perfume industry. Due to the high polarity of the solvent severe wrap-around can be observed. Wrap-around occurs when the second-dimension retention time exceeds the modulation time. Components then elute in subsequent second-dimension chromatograms and show up as spurious, broad peaks.

3.3.2 Data handling and pre-processing

After acquisition and integration in EZ-Chrom, the data were exported to Common Data Format (CDF) format and imported into the MATLAB environment using the NetCDF toolbox [104].

Integration

In-house developed MATLAB routines were used for demodulation of both the detector output and the retention times of integrated areas. The chromatographic data is visualized through a colour plot in greyscales. Superpositioned onto the colour plot are the peak apices to visualize the quantitative information. Summated areas are calculated through a polygon summation box and processed further in Excel. Figure 3.5 gives the shows an apex plot. The dots in the chromatogram indicate identified and quantified peaks by the integration software.

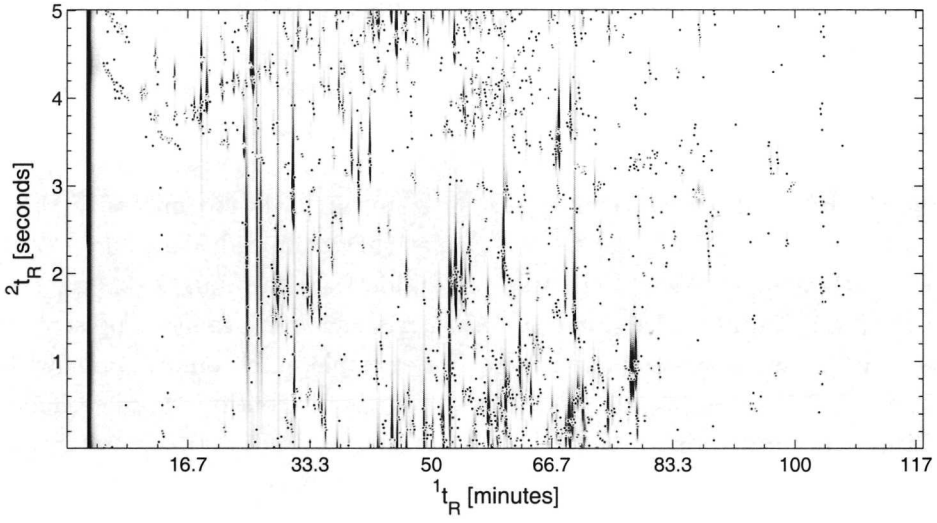


Figure 3.5: Apex plot of a typical perfume sample.

Peakfitting

Prior to the application of data analysis methods, data pre-processing is crucial. In this case the following steps were used:

Baseline removal: The offset, drift and wander of the baseline interfere with the quantitative information present in the chromatogram. Using a routine developed in-house, described in Section 4.2.2, page 59. The resulting baseline was subtracted from the original chromatogram. The baseline was calculated in such a way that no negative results in the baseline subtracted signal were produced.

Data stacking: Multiway methods require the data to actually be organized in a multiway orientation. Therefore, all GC×GC chromatograms are stacked on top of each other. The resulting matrix has the dimensions $(I \times J \times K)$ of $(1000 \times 250 \times 32)$.

Selection: Since in this study we are only interested in the concentration profiles of individual components, only the peaks of interest were selected. The typical selection window is 5 columns (first dimension) and 25 rows (second dimension) wide. The remaining (selected) matrix has typical dimensions of $(I \times J \times K)$ $(5 \times 25 \times 32)$. For each of the components of interest a separate sub-matrix was created.

Alignment: As in all chromatographic experiments, the actual retention times vary slightly from run to run due to small deviations in, for example, the temperature profile, the flow, the sample matrix and the (manual) injection. Shifted peaks are easily recognized by the human eye, because peak patterns remain identical. Thus, for user-supervised integration this is not a big issue. Data-analysis methods, however, are extremely sensitive towards shifts, and need a pre-processing step in order to minimize their effects. Bylund *et al.* [106] used Correlation Optimized Warping (COW) prior to Parafac analysis to eliminate retention-time shifts in LC-MS.

Elimination of shifts on a global scale, using all shift information present in the entire chromatogram, is preferred. For example, in chromatograms with a longer injection delay all peaks shift to higher retention times. Global shifting prevents individual peaks from being shifted to lower retention times. On a local scale the latter might occur, because no prior knowledge on shift profiles for individual peaks is present.

The observed shifts in this study are at most 4 points in the first dimension (20 seconds) and 20 points in the second dimension (0.4 seconds). The origin of these shifts is likely to be differences in the sample matrix, but also in operating conditions, which slightly differ from run-to-run. Synchronization (*i.e.* the simultaneous start of data acquisition and the GC run) is solved in the hardware.

Instead of solving all retention-time shifts (globally), we applied a correlation-optimized shifting based on the so-called inner product correlation [42] to the local selections. The inner-product correlation is defined as:

$$r_{(\boldsymbol{A},\boldsymbol{B})} = \frac{tr(\boldsymbol{A}^T\boldsymbol{B})}{\sqrt{tr(\boldsymbol{A}^T\boldsymbol{A}) \times tr(\boldsymbol{B}^T\boldsymbol{B})}} \tag{3.4}$$

Where:

$r_{(A,B)}$	Correlation coefficient between matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$.
$\boldsymbol{A}$	Standard matrix.
$\boldsymbol{B}$	Sample matrix.
tr	Trace function (sum of all diagonal elements).

A standard was used as reference and all other selections were aligned with this standard. By shifting the selection window over a predefined grid and simultaneously calculating the correlation, a best-fit position was found and stored. Restricting the permissible number of steps in the shifting process prevents the selection of a neighbouring peak belonging to a different component.

The actual calculations with the Parafac, Parafac2 and NPLS routines are simple and fast. Decomposition of the selected sub matrix (with the dimensions 5×25×32) with Parafac takes about 1 second calculation time. Parafac2, and to a lesser extent NPLS, take considerably more time, but still not exceeding half a minute. The model inputs are the peak selection (after shifting), the number of expected components and constraints for the calculation. Normally, a one component Parafac model is sufficient. However, if the captured variance is too low (<80%), an additional component can be introduced. If the resulting calibration line does not yield a physically realistic description, the additional component does not contribute to a better model.

3.4 Results

Conventionally, chromatograms are integrated in order to obtain quantitative data. Thus, in the context of quantitative chromatography, integration can be regarded as a benchmark technique. The results obtained with other, multiway methods, such as Parafac, Parafac2, and NPLS, should not differ from those obtained by integration.

3.4.1 Alignment

The most critical step in the use of mathematical models to describe chromatographic data is alignment. Two chromatographic axes, as encountered in GC×GC, make this problem even more challenging. A global shifting routine experiences great difficulties in dealing with 'wrap-around'. Therefore, we selected a window around a peak in the GC×GC chromatogram of the standard ('reference') sample and used it as template. The same selection window was used for the next injection ('sample') and between the two matrices an inner-product correlation was calculated. The selection window for the sample was shifted across the chromatogram two columns to the left and to the right and up to ten points up or down. For each shift the inner-product correlation was calculated (105 shift positions). The shift with the highest correlation was assumed to be the best alignment. The same procedure was repeated for all injections, standards as well as samples. An inspection of the chromatograms revealed that the correlation-based shifting was a good and fast method to eliminate shifts on a local scale.

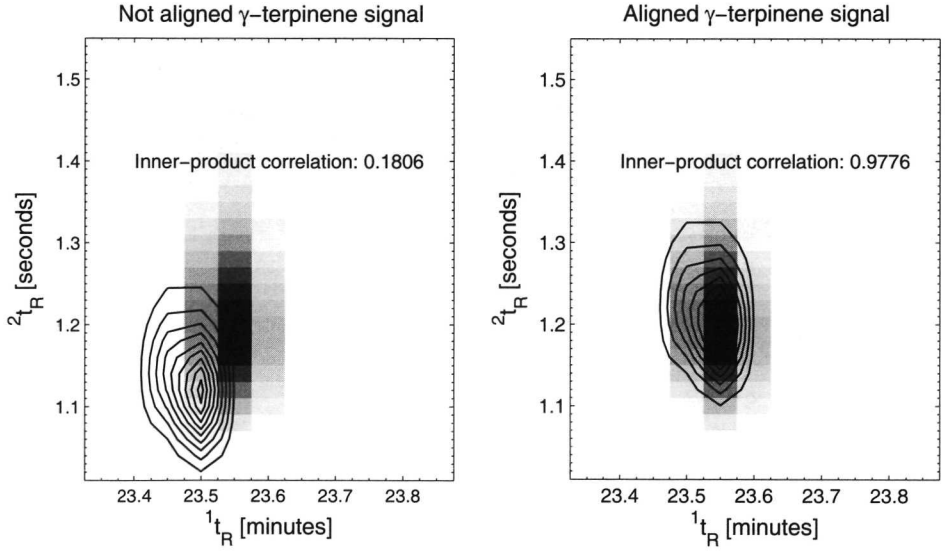


Figure 3.6: Effect of shifting (alignment) of a peak in a standard. Superpositioned on top of the chroma^{2}gram is a contour plot of a second chroma^{2}gram.

In this procedure no interpolation was involved and the automatic shifting

of 32 injections for a single component is completed in about 5-10 seconds. In Figure 3.6 the result of shifting was illustrated.

It should be emphasized that the improvement in correlation is not as dramatic in each sample as in the example of Figure 3.6. Samples containing low concentrations of the selected components yield lower correlation coefficients due to low signal-to-noise ratios (see Figure 3.7), but the highest value still corresponds to the best alignment. Even for samples containing other peaks in the immediate vicinity of the component of interest, shifting based on inner-product correlation appears to work properly.

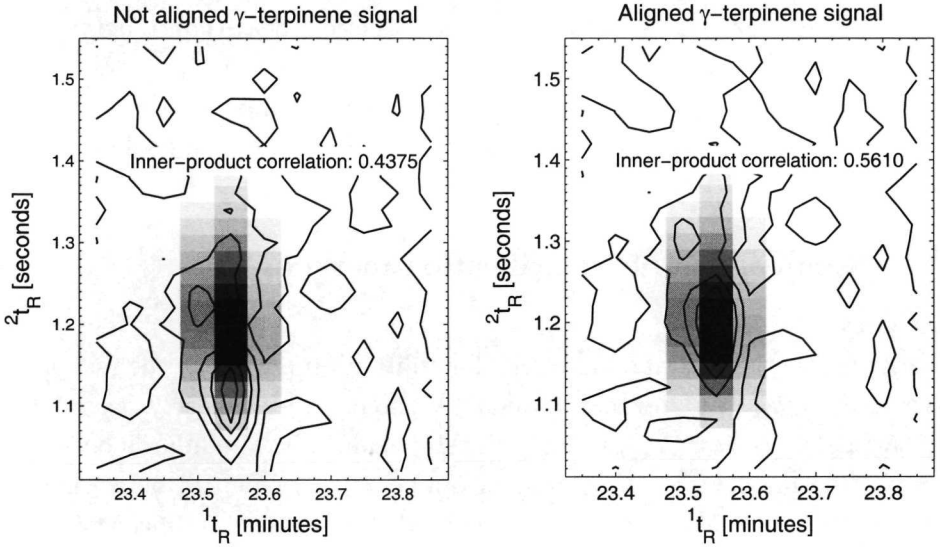


Figure 3.7: Result of shifting (aligning) performed on a peak in a sample.

After the alignment step the responses are calculated and corrected using the concentration and response of the internal-standard peak. In some samples, the selected local window contained more than one component. A theoretical advantage of the mathematical models described in the Theory section is the possibility of deconvolution, *i.e.* the reconstruction of pure-component elution profiles from overlapping peaks. The only condition is that the number of expected components is specified when applying the models. Overestimation of the number of components leads to an improved fit of the model, but the calculated factors (profiles) do not adequately describe the real factors.

Underestimation of the number of components also can lead to anomalies in the calculated peak profiles and responses. In the present samples and for the selected target analytes, a single component/factor model was sufficient to describe the variance in the local models. For samples containing two (or more) peaks in the selection window, additional factor(s) in the Parafac model can be considered. This should result in pure-component elution profiles for the target analyte and for the interfering component(s). However, if the additional peaks are found in only one or some of the samples, the introduction of additional factor(s) results in the modeling of the residuals of the first component. This is inherent to the least-squares criterion, which is used to minimize the residuals. The introduction of a second factor will always reduce the sum of squares, but it may lead to erroneous profiles and concentrations. The same aligned data are used as input for the different mathematical methods. Differences in calculated responses are solely originating from the methods.

3.4.2 Comparison of quantification methods

Linearity

In order to use the described methods for calibration purposes, the response (corrected using the internal standard) should vary linearly with the concentration. To test the linear relationship, calibration standards between 10 and 1500 mg/kg were measured in duplicate, interspersed between the samples. The correlation coefficient was used as a measure of linearity.

Correlation	Terpinene	Citronellyl	DMA	Lavandulyl	Eucalyptol	Mentone
Integration	0.9999	0.9997	0.9997	0.9996	0.9998	0.9997
Parafac	0.9979	0.9983	0.9988	0.9980	0.9973	0.9993
Parafac2	0.9987	0.9992	0.9989	0.9979	0.9976	0.9993
NPLS	0.9985	0.9986	0.9989	0.9972	0.9980	0.9993

Table 3.2: Correlation coefficients for all components with the various quantification methods.

Some differences in the correlation coefficients obtained using the three models are expected, since the ways in which the responses are calculated differ fundamentally due to constraints. In general, all methods revealed a good linearity (Table 3.2). It can be concluded that all methods result

in linear relationships between response and concentration. Integration performs best with respect to linearity.

Accuracy

A second calibration standard was measured as the last sample in this dataset under slightly different conditions (lower head pressure) to induce different peak shapes. This standard was treated as a sample and the concentrations were calculated for each component with integration, Parafac, Parafac2, and NPLS. Ideally, the calculated concentrations should be identical to of the true values. A deviation of 5% was thought to be acceptable.

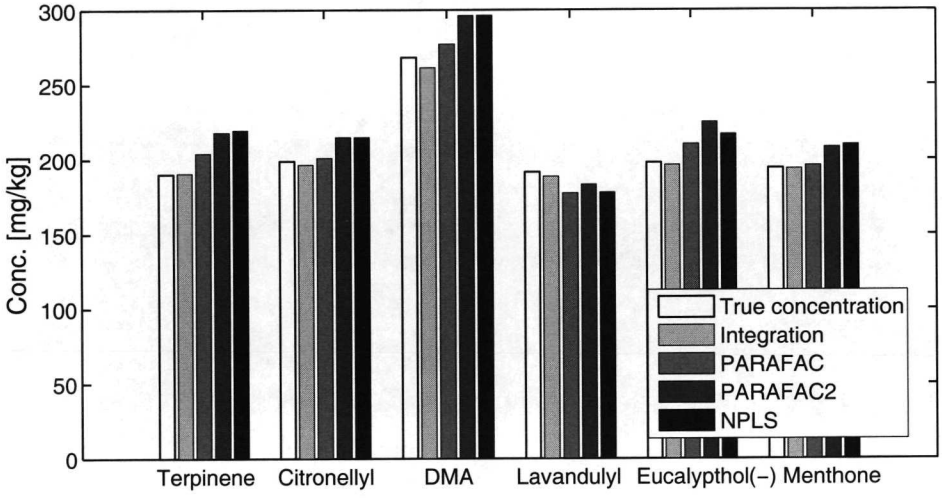


Figure 3.8: Accuracy of various methods based on the analysis of a reference mixture with known analyte concentrations.

As can be seen in Figure 3.8, integration performs best for (almost) all components. Parafac2 and NPLS tend to overestimate the concentrations. Parafac is the most accurate of the multiway methods in the present case. The influence of the peak shape seems to be more detrimental for Parafac2 than for Parafac. This result is surprising, since Parafac2 should theoretically be capable of dealing with shifted peaks.

Calculated concentrations

The results for the four samples, six target compounds and four quantification methods are given in Table 3.3.

Sample	Method	Terpinene	Citronellyl	DMA	Lavendulyl[a]	Eucalyptol	Mentone
M2	Integration	1830	405	16	58100	**800**	160
	Parafac	1880	405	40	55000	**310**	150
	Parafac2	1890	406	114	54200	**480**	157
	NPLS	1900	407	40	53300	**296**	150
M4	Integration	2.2	3.8	**100**	123000	16	36
	Parafac	4.3	6.8	**44**	115000	20	32
	Parafac2	6.2	11.8	**54**	118000	23	33
	NPLS	4.3	6.8	**44**	109000	21	32
M6	Integration	480	30	154	30300	**2790**	22
	Parafac	480	34	170	31000	**1330**	19
	Parafac2	498	36	254	29900	**1560**	22
	NPLS	491	34	172	29700	**1330**	19

[a] In real samples the peak of lavandulyl acetate is perfectly co-eluting with ortho-tertiary butyl cyclohexylacetate (OTBCA) present in concentrations up to 30% [*w*/*w*] in the sample. Both components have similar retention indices in both separation directions and completely overlap, even in GC×GC.

Table 3.3: Concentrations [mg/kg] in real samples obtained using integration and using the multiway methods. Bold numbers indicate large deviations.

In four cases there is a major difference between the methods (DMA/Sample4, Eucalypthol/Sample2 and Eucalypthol/Sample6, indicated in bold). These differences most likely originate from the shift routine, since the differences between the three multiway methods mutually are much smaller than those between the multiway methods and integration. Especially at low concentrations (<10 mg/kg), multiway methods systematically overestimate (assuming that integration provides the correct answer!). This might be due to the baseline removal, which does not allow negative baseline values. The result is a minor offset in the baseline, which can lead to overestimation at low concentrations. No experiments were performed to verify this (*e.g.* via standard addition). Surprisingly, the highest concentrations in almost all cases are found with Parafac2.

Limit of quantification(LOQ)

The limit of detection in GC×GC is primarily determined by the signal-to-

noise ratio of the peaks detected by the FID. The LOQ generally is defined as three times the S/N ratio and would obviously be identical in all four cases. Quantification, however, is also affected by the ability to differentiate between signal and noise. This is where integration and peak fitting approaches differ. In the case of integration, the minimum-area setting results in limits of quantification between 3 and 10 mg/kg, depending on the component of interest (purity, FID response factor). In the case of Parafac, Parafac2 and NPLS, the minimum detectable amount is less easy to determine, since it is also influenced by other samples in the dataset. If, for instance, the dataset is constructed solely from samples with low concentrations, then the minimum limit of quantification is expected to be lower then in case of a set of highly concentrated samples with only one dilute one. In this case, we estimate the limits of quantification for the multiway methods to be in the range of 6 to 20 mg/kg, somewhat higher than those obtained with integration.

Comparison of integration and multiway methods

The logarithmic scale forces the attention on the low concentration part of the comparison, where the largest deviations appear.

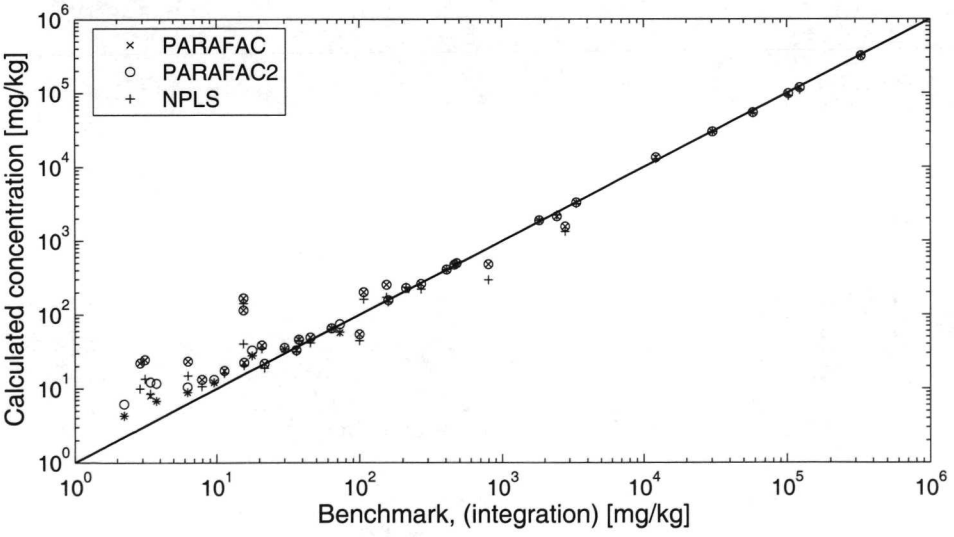


Figure 3.9: Comparison of quantification methods with to integration (regarded as benchmark technique).

On a logarithmic scale the results obtained with integration and with Parafac show a linear relationship without any real inconsistencies (Figure 3.9). The observed differences mainly appear in the low concentration region, near or below the LOQ.

Precision

One may expect that multiway methods yield a lower precision than conventional integration. This is probably true for simple (gas) chromatograms containing only a limited number of peaks, but in this particular case it turns out that precision is comparable, if relative standard deviations (r.s.d.) are used. In Figure 3.10, the r.s.d. for triplicates are shown as function of the calculated concentration. It appears that the three multiway methods do not show substantially higher r.s.d.'s then does integration. Differences appear in the low concentration region (<10 mg/kg), where the multiway methods are expected to perform worse. On average, multiway methods do not perform significantly worse than integration with respect to precision.

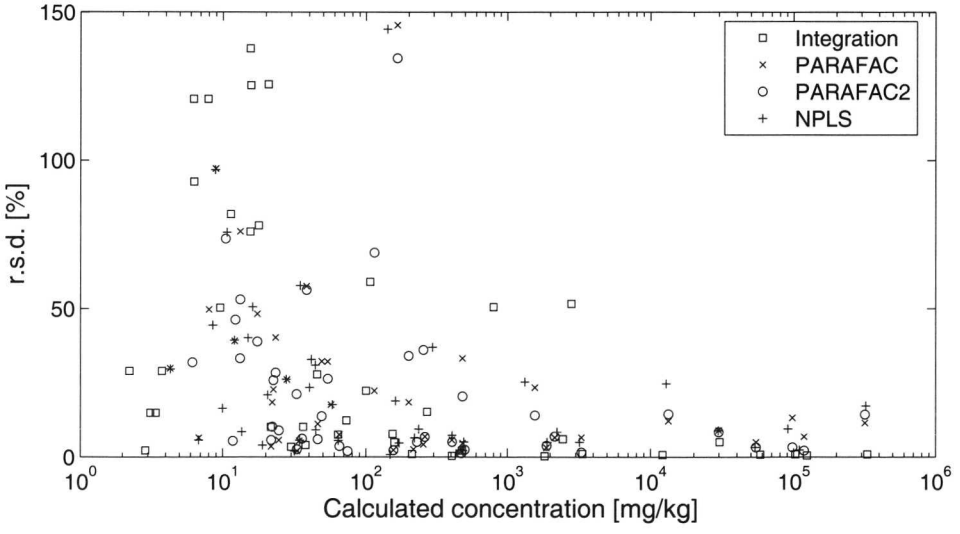


Figure 3.10: Errors (r.s.d.)obtained by various methods as function of concentration for seven target analytes.

Speed

The rigorous quantification of large GC×GC datasets with integration is a very time-consuming exercise. It requires about two minutes per component per chromatogram to integrate (GC×GC) slices, due to the manual combination of peaks. For the present dataset of 32 injections and 13 components, 13 hours of analyst effort were required to integrate all peaks. Further processing with Excel takes another three hours. This could be improved by the use of routines that combine the successive apices. However, this would lead to large result tables containing all the combined slices. From these, a selection has to be made of components of interest. The quantification by Parafac or NPLS takes only two minutes per component, regardless of the number of chromatograms. In the present study, 30 minutes proved sufficient to fully quantify all the target components in all the chromatograms. Further processing in Excel is easier (about 1.5 hours), since Parafac and NPLS yield an array of concentrations that can be directly imported. In total, integration takes about 16 hours, whereas Parafac and NPLS require about two hours for the total set.

3.5 Conclusions

Integration is the preferred method for accurately determining concentrations in GC×GC. This method is, however, very time-consuming and labour-intensive. Multiway methods, such as Parallel Factor (Parafac) analysis, its extension Parafac2, and multi-linear Partial Least Squares (NPLS), are all capable of estimating concentrations in the chromatograms. Especially constrained Parafac yields concentrations comparable to integration in terms of accuracy and precision. Due to different approaches in the multiway methods, a dramatic increase in productivity is found. Integration requires about 16 hours for the quantification of 13 components in 32 chromatograms, whereas Parafac and NPLS require only 2 hours. This aspect becomes increasingly important in the context of new GC×GC instruments equipped with jet-modulators and auto-injectors. The jet modulators permit higher data-acquisition rates (at least 100 Hz) and have the potential of increased numbers of peaks, while auto-injector units allow large numbers of analyses to give rise to large datasets. The shifting

routine developed for the multiway approach seems to work satisfactory on the dataset described in this Chapter. However, more experience is required to arrive at more definitive conclusions. It is also found in the present study that Parafac2 and, to a lesser extent, NPLS overestimate concentrations in comparison with integration. For NPLS this can be partly explained by the fact that the method calibrates using pure-component chromatograms, but predicts on multi-component samples. For Parafac2, however, this comes as a surprise, since the method was thought to be able to deal with retention-time shifts encountered in the first-dimension chromatograms, due to the inner-product structure. One of the reasons for this may be the fact that peaks in the first dimension are not shifted, but show a different peak profile, which is referred to in literature as "in-phase" and "out-of-phase" modulation [87]. This phenomenon leads to differences in the inner-structure property, but would only partially explain the systematic overestimation of the concentrations obtained by this method.

Acknowledgments

The authors would like to acknowledge Shell International Chemicals, specifically Jan Blomberg and Marcel van Duyn for their contributions to this Chapter.

Chapter 4

Chemometric group-type tools

Comprehensive two-dimensional gas chromatography (GC×GC) is now being used for a large number of applications. Three generic types of applications can be distinguished. One of these, group-type analysis, focuses on the quantification of groups of components. Often the components within such groups have common structural features and physical-chemical properties, or they exhibit similar behaviour with respect to legislation, health, or product specifications. If the separation mechanisms in the two dimensions match the most important differences between the different component groups, then GC×GC can achieve structured separations. However, to obtain accurate quantitative information on component groups several data preprocessing steps are necessary.

In this Chapter we describe the steps required to convert the signal obtained from the detector in GC×GC and the quantitative information obtained from the chromatography data system (*e.g.* peak areas) to a data matrix that can be analyzed by multivariate techniques. In addition, tools such as baseline correction and splining will be discussed. It will be described how quantitative data on component groups can be obtained. Finally, retention-time shifts are unavoidable in separation techniques. We describe a way to deal with such shifts in group-type analysis.

4.1 Introduction

Many natural and industrial products contain huge numbers of individual components. Conventional, one-dimensional chromatographic separations are often inadequate for answering questions on the composition of such very complex mixtures, due to peak overlap. Although comprehensive two-dimensional separation techniques offer a great increase in separation capacity, even these methods cannot provide the separation power required for a complete separation. At the same time, there is an increasing interest in the adequate characterization of both natural and industrial products. This trend is partly driven by process optimization, as well as by the need to assess product properties. Estimation of the negative (pollution index or general toxicity) or positive value (*e.g.* fuel quality for airplanes, cars and heating) does not usually require a complete separation.

This issue has been addressed in a classification scheme for chromatographic separations (Chapter 2 of this thesis). Three generic applications were identified, that require common strategies to translate chromatographic data into relevant information on the sample. The first of these generic applications was referred to as Target-Compound Analysis. This concerns the quantitative analysis of a limited number of predefined (targeted) components. Target analytes may, for example, be related to product specifications, process control or legislation. Separation of the components of interest from each other and from the matrix is generally required for this type of application. The second type of application, Group-Type Analysis, focuses on the quantification of component groups. For situations in which there is a strong correlation between the total amounts of certain component classes and product properties, such as toxicity, the separation between component groups is important, while separation within component groups is often undesirable. The last type of application, Fingerprinting, concerns the correlation between sample properties and chemical composition of a sample (*e.g.* oil spills). In certain cases, there is no *a-priori* information which component(s), component groups, or component profiles are related to a certain property. By considering the complete chromatogram as a fingerprint of the sample, correlations between properties and chemical composition can be established. These applications require a high resolution and a very high retention-time and response stability. One of the advantages of this classification scheme is

that commonalities in the data-processing strategies can be deduced. Each specific application type puts certain demands on both the chromatographic system and the data processing.

In this Chapter, we will focus on the second type of application, *i.e.* Group-Type Analysis. By far the greatest asset of comprehensive two-dimensional separations for the separation of group of components is the possibility to obtain structured chromatograms, which substantially aids component classification and identification [65].

The appearance of structured chromatograms is closely related to the concept of sample-dimensionality, introduced by Giddings [66]. It is based on the intrinsic properties of analytical samples. The sample-dimensionality is defined as the number of independent variables required to identify the components in a sample. It determines the susceptibility towards multi-dimensional separation techniques. Matching the separation dimensions to the sample dimensions will result in structured chromatograms. For example, if the second dimension separates strictly according to one of the relevant sample dimensions, parallel bands of component groups are formed along the first-dimension axis. Specific advantages of comprehensive two-dimensional gas chromatography (GC×GC) are the tuneable selectivity and the decoupling of volatility and polarity contributions to the analyte retention [65].

Examples of group-type separations by GC×GC have first been demonstrated for oil-product analysis [64]. Oil products typically contain overwhelming numbers of components, yet a limited number of chemical classes. Other examples of group-type analyses are the classification of fatty acids according to the degree of (un)saturation [68, 69] and the classification of PCB's according to planarity [70].

Obviously, the separation of highly complex samples into component groups puts requirements on the separation system. The column combination should provide highly selective separations according to the most-relevant sample dimensions. In this way, component groups can be separated from each other and from the matrix. The peak capacity of the separation system is of much less importance. The primary aim is the ability to quantify groups of components, rather than individual species. Therefore, ideally, the detector used for group-type separations should offer equal response for all members of a component group. For this latter reason mass spectrometry is not often used for these kinds of applications, other than for qualitative purposes (*i.e.*

establishing and validating the locations of classes in the chromatograms). Most quantification strategies involve the determination of peak areas and concentrations at the component level. This is useful for the quantification of a limited number of targeted components. However, for the complete characterization of very complex materials, such an approach would result in large amounts of (unreliable) data, *viz.* the retention times and peak areas of all individual chemical components. Clearly, this is not desirable. The ordered separation obtained by a properly configured and tuned GC×GC system should be exploited for the quantification of well-defined groups of analytes ("pseudo-components").

4.2 Theory

4.2.1 Comprehensive two-dimensional gas chromatography

Comprehensive two-dimensional gas chromatography (GC×GC) has been one of the most significant advances in the last decade for the characterization of complex volatile mixtures. This technique was pioneered and advocated by the late John Phillips [1–3]. A GC×GC system utilizes two different columns. The fist-dimension column is (usually) a conventional capillary GC column, with a typical internal diameter of 250 μm. Most commonly, this column contains a non-polar stationary phase, so that it separates components largely based on their vapour pressures (boiling points). The second-dimension column is considerably smaller (smaller diameter, shorter length) than the first-dimension column, so that separations in the second dimension are essentially much faster. The stationary phase is selected such that this column separates on properties other than volatility, such as molecular shape or polarity. The two columns are coupled using a so-called modulator. This device facilitates the continuous accumulation, refocusing and injection of small portions of the first-column effluent into the second-dimension column. With each modulation, a new second-dimension chromatogram is started. The technique can be called comprehensive if each chromatographic peak in the first dimension is divided into three or four fractions or "slices" (second dimension chromatograms), preserving the chromatographic information from the first dimension as much as possible [87].

The detector, which is positioned at the end of the second-dimension column, records the fast second-dimension chromatograms. At the end of a chromatographic run, the recorded data file contains many of these fast separations in series.

GC×GC has found its way to a large variety of application areas, such as oil and petrochemical products [64, 77, 107], fatty acids [68, 69], essential oils [60, 108] and atmospheric air [109]. The technique has also been the subject of a number of review articles [3, 6–9]. Pre-processing As mentioned before, the classification of chromatographic applications into three generic application types allows us to develop pre-processing strategies for each type of application. However, some general pre-processing steps are necessary for all GC×GC data, *i.e.* demodulation (or matricizing) and baseline correction.

Demodulation

Most gas chromatographs and certainly most GC×GC instruments are controlled by chromatography data systems (CDS). In most cases, the CDS also is used to collect the data.

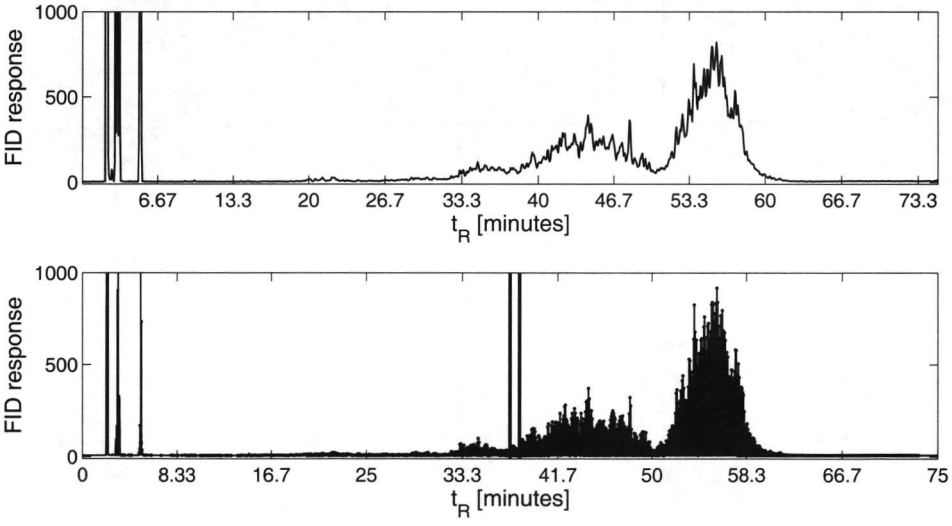


Figure 4.1: 1D reconstruction (upper chromatogram,(a)) and modulated signal (lower chromatogram, (*b*)) obtained for a Tridecanol sample. Vertical lines in the lower chromatogram indicate selection of Figure 4.2 and Figure 4.3.

As described before, the "linear" signal obtained from the instrument is a long series of fast chromatograms. Transforming the data into a matrix format ("matricizing") is generally referred to as demodulation [5].
The time-intensity signal of a GC×GC instrument can be analyzed with conventional (and well-established) integration routines. By integrating the linear signal, components at the 'edges' of the chroma^{2}gram are correctly quantified. A minor disadvantage of this approach may be the continuous (saw-tooth) variation in peak width within the second-dimension chromatograms. This makes it difficult to estimate correct integration parameters. However, in practice this has not proven to be a serious issue.
In Figure 4.1, the 1D-reconstructed signal of a heavily branched C_{13} alcohol (derivatized with N-Methyl-N-(trimethylsilyl)trifluoroacetamide, MSTFA) is presented. In this chromatogram unsatisfactory separation in D1 is achieved. Figure 4.1b shows the modulated 1D signal. If we zoom in at the selected region of the signal, the individual second-dimension chromatograms can be observed (Figure 4.2).

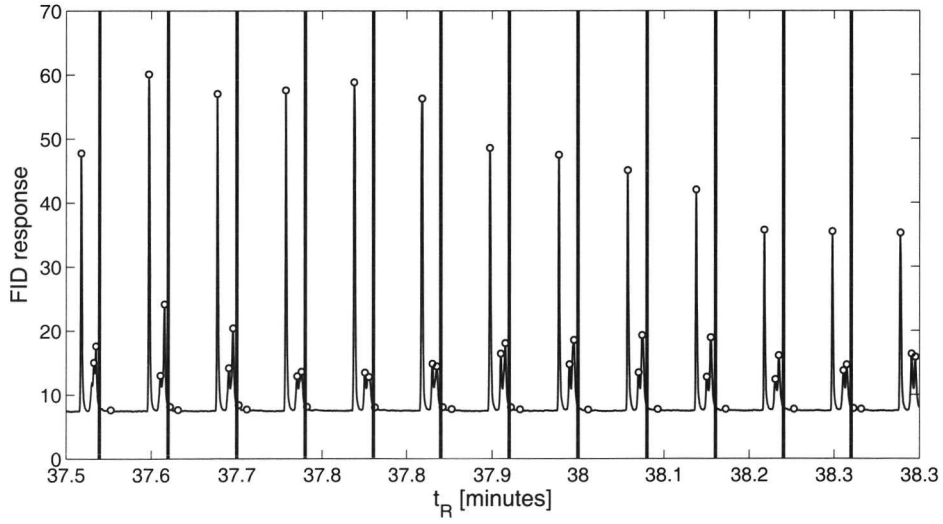


Figure 4.2: Linear chromatographic trace and modulation sequence.

Vertical lines in Figure 4.2 indicate the modulation period, which in this case was four seconds.

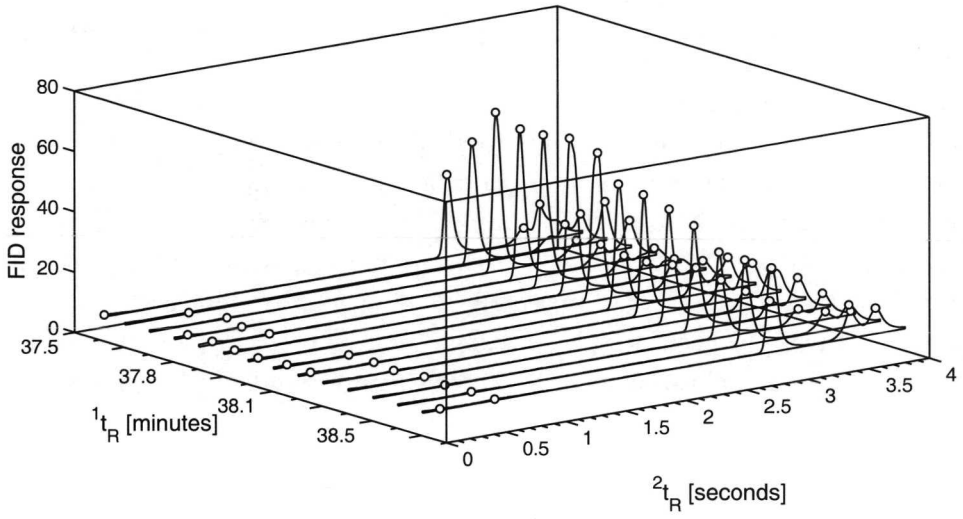


Figure 4.3: Demodulated segment of Figure 4.2 selection as shown in Figure 4.1).

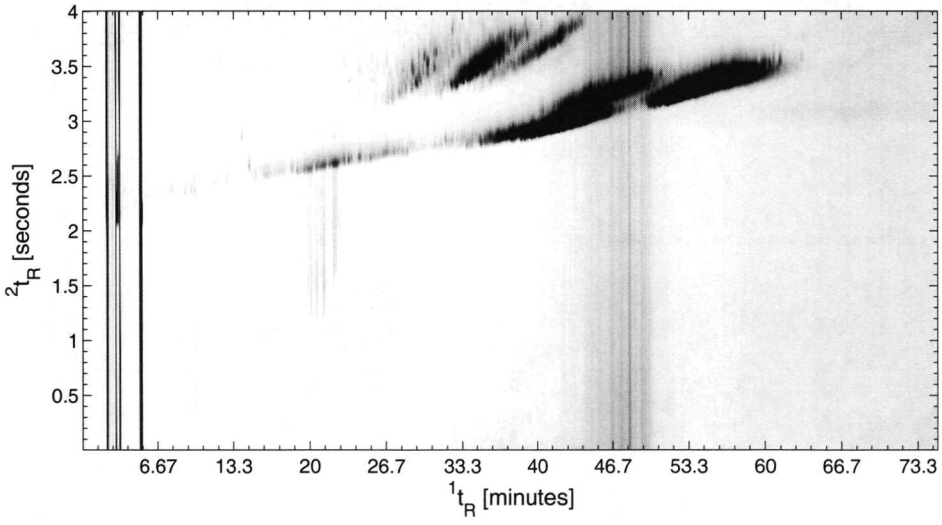


Figure 4.4: Demodulated, two-dimensional chromatogram (or "chroma^{2}gram ") of Tridecanol sample (same data as Figure 4.1*b*).

Note that the modulation period can be clearly discerned in the chromatogram, but that the exact time of injection in the second-dimension column is difficult to determine.
All vertical lines can be shifted to the left or to the right simultaneously. In a chroma^{2}gram this corresponds to all peaks moving up or down. This implies that the absolute vertical position of the component bands in a chroma^{2}gram is not known.Demodulation involves cutting the linear signal into individual second-dimension chromatograms. Rotating all individual chromatograms over 90^0 and projecting them in a three-dimensional space, results in the typical contour plot or colour plot. In a colour plot, each point in the detector signal is projected by a colour, corresponding to the intensity. In a contour plot, lines of equal intensity are calculated. Each contour line can be considered as a slice of the two-dimensional gas chromatogram at a certain signal height. To avoid confusion with (two-dimensional) chromatograms obtained from one-dimensional separations, (three-dimensional) chromatograms obtained from two-dimensional chromatograms are referred to as chroma^{2}grams.

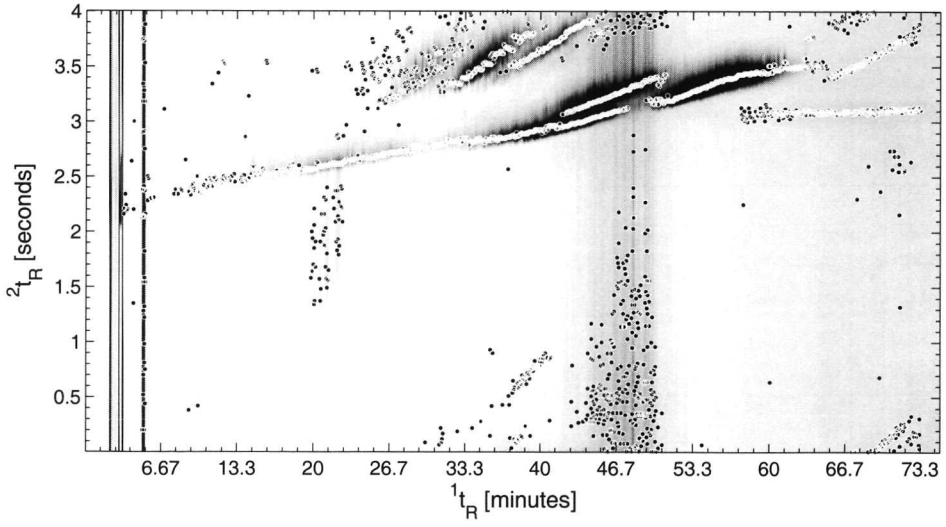


Figure 4.5: Demodulated chroma^{2}gram of 4.4, with peak-apices.

Figure 4.3 shows the selection of Figure 4.2 after the demodulation step in the form of a so-called "waterfall plot". In this representation the

second-dimension chromatograms are displayed as discrete lines, with no information between two consecutive second-dimension chromatograms. Demodulation of the complete chromatogram results in a "black-and-white" or grey-scale colour plot (b/wC) (Figure 4.4). The quantitative information (*i.e.* peak apices) obtained from the CDS requires a demodulation step as well. Each individual peak retention time is converted into a set of retention coordinates. The ${}^1t_R$, is obtained by the maximum integer number of times the modulation time 'fits' in the retention time. The remainder of the retention time is ${}^2t_R$. Plotting these coordinates onto the chroma^{2}gram results in Figure 4.5.

4.2.2 Baseline correction

An important data pre-processing step is the correction for baseline drift in the chroma^{2}gram. Especially for components present in low concentrations this aids in the visualization of the peaks. The detector (background) signal tends to increase at higher temperatures (Figure 4.6).

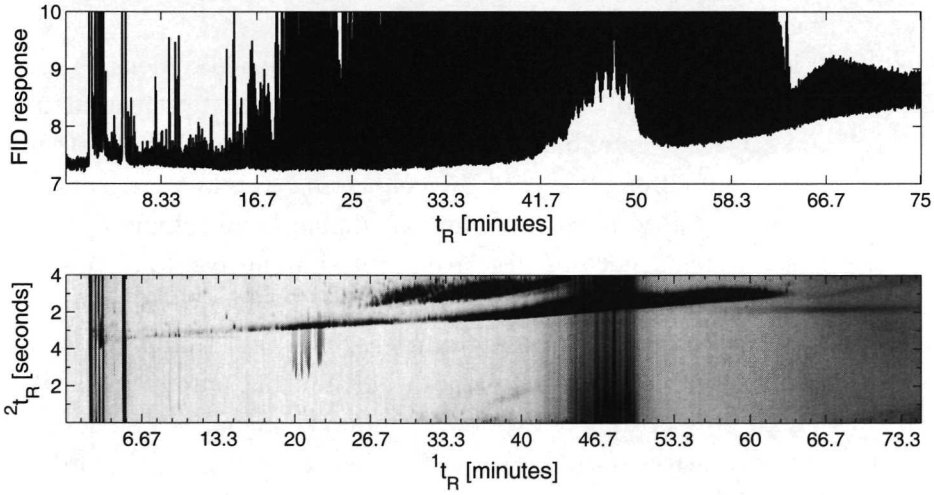


Figure 4.6: The upper (modulated) chromatogram shows the Tridecanol signal. Clearly visible is the increased baseline drift with increased oven temperature. The lower (demodulated) chromatogram shows the effect of baseline drift in the chroma^{2}gram.

This is equally the case in conventional one-dimensional GC and in GC×GC and it is caused by increased bleeding of the stationary phase from the column. Several methods for baseline correction can be distinguished. A modern GC allows a blank signal to be recorded. This signal is subtracted from each following chromatogram. Whereas this seems to be a convenient method, performing this step typically destroys the original signal. Any variations in the blank run will affect each following chromatogram. An alternative is to use software tools for estimating the 'real' baseline level of chromatograms. For GC×GC, a statistical approach has been developed by Reichenbach *et al.* [110]. Use is made of the region in a chroma^{2}gram in which no components elute. This may be a valid approach for oil samples, in which paraffins are the least-polar components and the polarity range of the components is limited. However, for applications in which an empty region is absent, this statistical approach will lead to erroneous results. This will especially be the case if the baseline-corrected data are used to extract quantitative information. For these reasons, we developed a baseline-correction tool, which estimates the baseline under a 2D-chromatographic signal. The assumption is that there is always some baseline present in the individual second-dimension chromatograms. Our algorithm detects these points, which can be considered as "knots".

Since there is a connection between the start of a second-dimension chromatogram and the end of the previous chromatogram, application in a chroma^{2}gram can result in anomalies. Therefore we project these knots into the one-dimensional chromatogram. By 'connecting' the knots results in a baseline reconstruction from the original, one-dimensional signal. A second step prevents the occurrence of negative points on the resulting baseline. Subtraction of the reconstructed baseline from the one-dimensional trace results in baseline corrected chroma^{2}grams (Figure 4.7). Compared to the original chromatogram of Figure 4.6 (both on the same intensity scale), a clearly enhanced presentation of the data is obtained. Especially signal components of low intensity can more easily be discerned. During the process, the quantitative information entailed in the data remains completely intact.

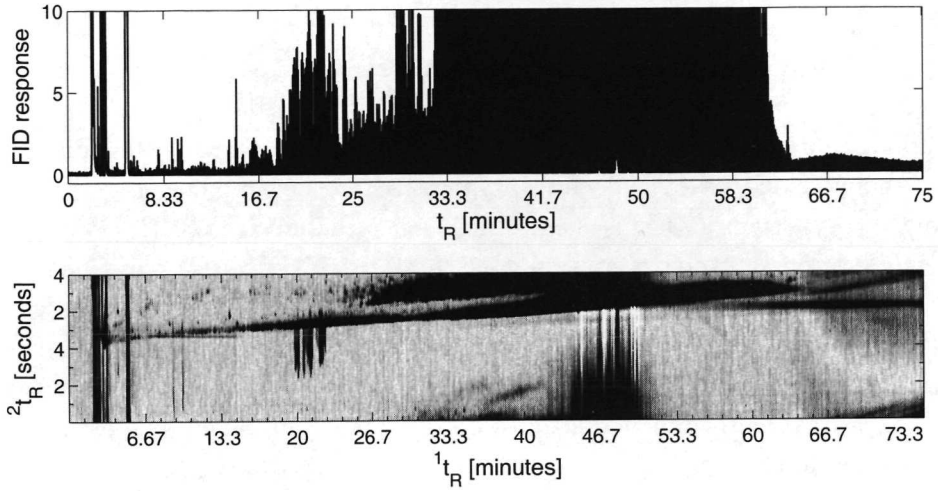


Figure 4.7: The effect of the baseline correction on the modulated (upper) and demodulated chromatograms.

4.2.3 Splining

The position of the various component groups in the chroma^{2}grams of Tridecanol (Figure 4.4) may easily give rise to some confusion. Firstly, it is not clear where the least-polar components elute, because the vertical anchoring of the picture is arbitrary. In addition, this chroma^{2}gram seems to contain so-called "wrap-around". This occurs when the second-dimension retention times exceed the modulation time, so that components show-up in subsequent second-dimension chromatograms. Typically, wrap-around manifests itself in the form of broad peaks, which often overlap with low-polarity components from subsequent second-dimension chromatograms. This is not the case in Figure 4.4. The wrap-around in this Figure rather seems to result from an incorrect presentation of the data. With increasing first-dimension retention times, the second-dimension times also appear to increase. This results in a tilted orientation of the component groups.
One explanation for this phenomenon may be timing errors, either in the modulation or in the data-acquisition. However, this does not concur with the observation that the effect increases when longer second-dimension columns are used. A more likely explanation is provided by the variation

of the flow rate over time. The chromatograms of Figure 4.1-4.8 were collected in the constant-pressure mode. With increased time (*i.e.* increasing temperature) the viscosity of helium increases, resulting in a decrease in the linear gas velocity. As a result, the dead-time (t_0) in both dimensions will gradually increase and the increase in ${}^2t_0$ is reflected in Figure 4.8. Members of a homologous series will elute at increased second-dimension retention times (${}^2t_{R,i}$), even if their second-dimension retention factors (${}^2k_i$) remain constant. Constant-flow operation may overcome this effect. However, a constant flow regime is very difficult to maintain in a two-dimensional system, with different columns (of different diameter) and (modulation) capillaries connected in series. However, the strong slanting in Figure 4.8 seriously complicates the assignment and interpretation of component groups in the chromatogram.

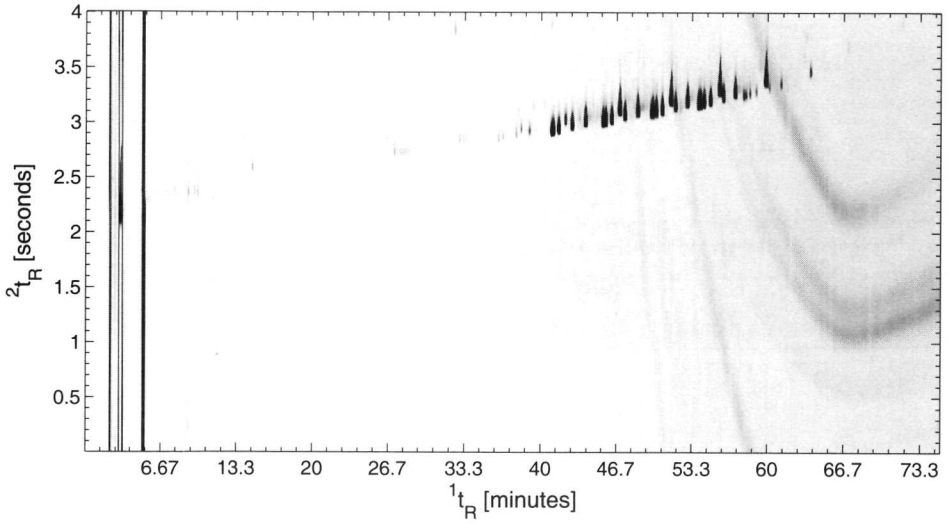


Figure 4.8: Chroma^{2}gram of Lialette (slightly branched C_{12}-C_{15} alcohol, derivatized with MSTFA after demodulation.

Correcting the chroma^{2}gram therefore seems necessary for the accurate quantification of component groups. Compensation for the tilted appearance of homologous series can be achieved by applying variable selection to the data matrix, such that constant (low) second-dimension retention times are imposed on the least-polar components. Conceptually, this can be seen as a

correction for the increase in the dead-time. The correction process is referred to as 'splining'. Phillips incorporated such an algorithm in one of his first data-analysis programs.
It must be emphasized that the main objective of splining is to improve the visual representation of the data. The quantitative information (*i.e.* the peak areas) is not affected. The first step in the splining process is to identify a series of homologous in the chroma^{2}gram and to establish their positions. In the case of oil samples, this is rather straightforward. The normal alkanes often stand out in the chroma^{2}grams, because of their relatively high abundance. If not, a mixture of *n*-alkanes can be injected separately. In either case the positions of the *n*-alkanes can be unambiguously assigned. It is important to correctly assign all members throughout the entire chromatogram. Between the selected points linear interpolation is performed.

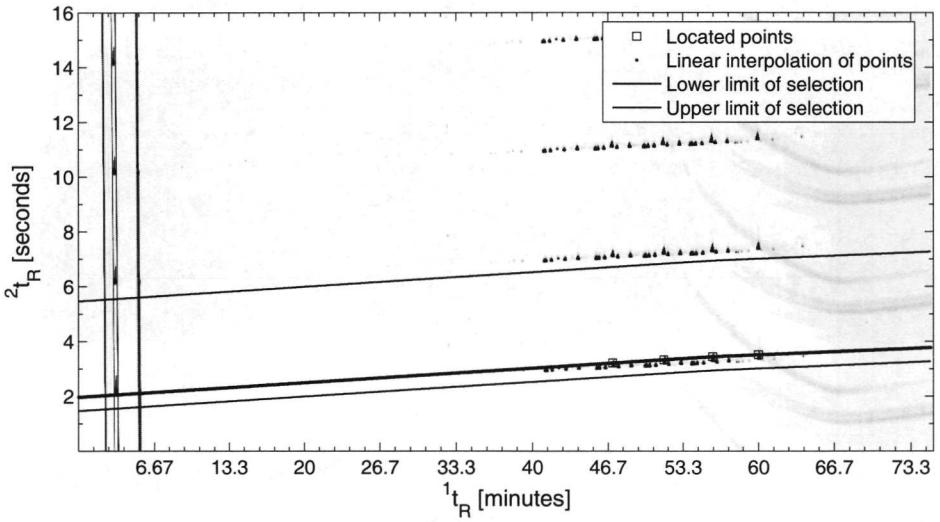


Figure 4.9: Illustration of the splining process.

Towards the beginning of the chromatogram, linear extrapolation from the first two selected points is used to calculate the locations of absent *n*-alkanes. Towards the end of the chromatogram, linear extrapolation from the last two selected points is used. In many cases, the chroma^{2}gram from which the splining function is determined is not the actual sample, but a separate injection of a series of homologuous. An evident requirement in the splin-

ing process is that the "real" samples and the sample from which the spline is calculated are measured under identical conditions on the same column set. In our practice, we insert the synthetic mixture for establishing the spline function in between the samples. This minimizes possible run-to-run variations. The set of homologue locations is applied to the original data matrix. Since the selection can easily exceed the boundaries of the second-dimension retention axes, a second and third matrix are added to the original one. Graphically, several chroma^{2}grams are super-positioned on top of each other (Figure 4.9). The matrices are aligned in such a way that individual (second-dimension) chromatograms within the initial matrix are continued in the attached matrices. Since the selected points are preferably located at the bottom of the chromatogram, but not at the very bottom, a small offset is incorporated in the splining function. From the selected points, the offset is subtracted and the final selection for that column starts there and ends a modulation time further (lower and upper selection limits, respectively). In Figure 4.9 this process is visualized.

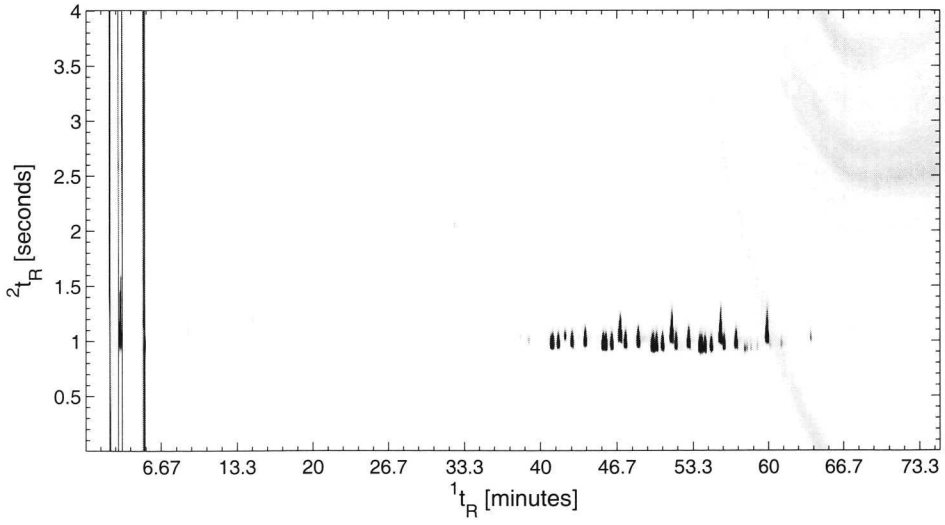


Figure 4.10: chroma^{2}gram of Figure 4.8 after applying the splining procedure.

The squares indicate peak positions of homologous in the sample. In this case, MSTFA-derivatives of linear alcohols from C_{12} up to C_{15} were used.

Between these four points, linear interpolation was used. Extrapolation towards lower and higher first-dimension retention times is based on the first and last two selected points. The black line in Figure 4.9 indicates the interpolated and extrapolated positions. These points will be positioned at a small offset near the bottom of the 2D chromatogram. The lower selection limit in Figure 4.9 is established by subtracting the offset from the black line. The upper limit is calculated by simply adding the modulation time. The result of the splining process is shown in Figure 4.10. The homologous series appears with constant second-dimension retention times. Application of the same selection to the Tridecanol sample of Figure 4.5 results in Figure 4.11. In this chromatogram, the apparent wrap-around is removed. The peak coordinates in the peak (quantitative data) must also be corrected. To do so, each peak coordinate is corrected for the offset from Figure 4.9. Negative second-dimension retention times are avoided by placing the peak in the previous column (previous slice) and adding the modulation time.

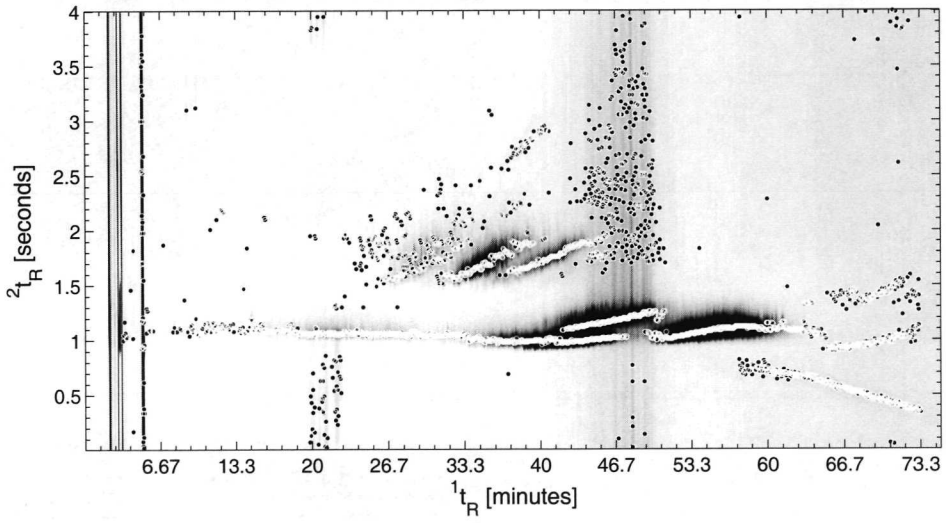


Figure 4.11: chroma^{2}gram of the Tridecanol sample (Figure 4.4) after applying the splining procedure.

4.3 Quantification

Although in the first few years of its existence GC×GC was mainly was used for qualitative analysis, the technique provides excellent quantitative data. Beens *et al.* [79] have demonstrated the quantitative performance by comparing the results obtained by one-dimensional GC with those obtained by GC×GC on an identical sample. However, the usefulness of GC×GC for quantitative analysis greatly depends on the availability of software.
Image-processing tools have been applied to extract quantitative information from two-dimensional chromatograms [111]. This approach is somewhat controversial, since the direct result of a two-dimensional separation is a linear chromatographic signal. Moreover, the image-processing approach requires a baseline-corrected signal. On the other hand, integration of peaks in conventional chromatograms has been applied to obtain quantitative chromatographic during many years. Integration procedures have been gradually improved and optimized and they can now be considered as highly robust and reliable. Since almost all GC×GC systems are controlled by some sort of

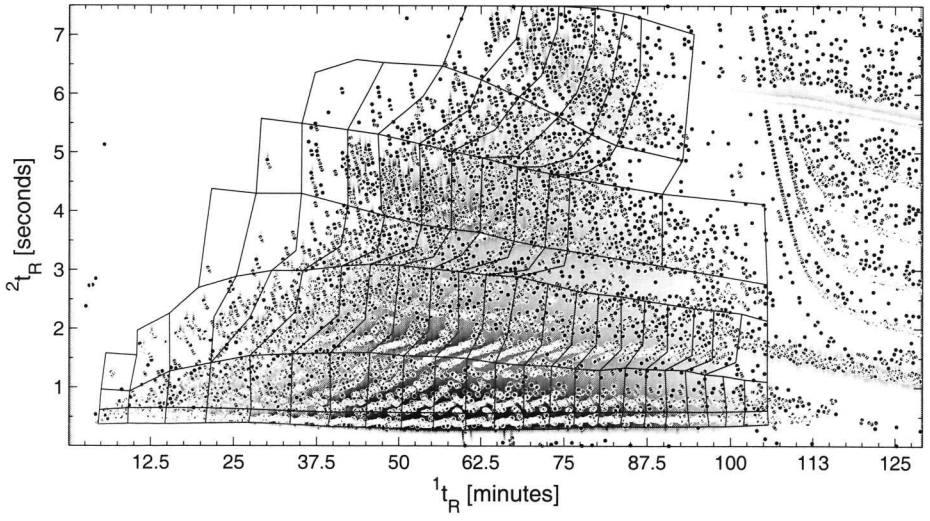


Figure 4.12: Two-dimensional chromatogram of a diesel sample (recorded at 250 kPa inlet pressure) in which 104 component groups are indicated by polygons.

chromatography data system (CDS) and since data are recorded in the form

of a series of conventional chromatograms, integration seems a very logical data-processing step. Since the modulation process yields several slices across each first-dimension peak (at least three to four to call the method comprehensive [87]), each individual component peak gives rise to (at least) three to four individual second-dimension chromatograms. The peak area and retention time representing the peak are shown as a peak apex on top of the chroma^{2}gram.

When the focus is on target-component analysis (Type-I application), these different peaks with different apices must be combined for the specific analytes. However, for group-type (Type-II) applications, the focus is on the quantification of groups ("pseudo-components") rather than on individual components and combining the apices into analyte peaks is not required.

Quantification of pseudo-components can be achieved by selecting component groups of interest. These groups are selected by marking their boundaries in the chroma^{2}gram. A number of demarcation points can be selected for this purpose. In our software, the number of points that mark the boundaries of a group is not limited, so that an endless variety of shapes can be formed. A convenient feature is that the positions of the selection polygon can be given "magnetic" properties. Each new point within a certain predefined radius of an already defined polygon will automatically be drawn to a previous position. In this way, a mesh-grid without any gaps can be placed on top of the chromatogram (Figure 4.12). A summation of all areas of peaks that have their apices within a polygon selection results in the total peak area for the component group. Manual action is required to draw a polygon around a component group of interest. This makes the construction of a complete quantification template a rather time-consuming exercise. However, automation is very difficult and would require some sort of image-processing approach.

4.3.1 Retention-time shifts

Processing of a series of chroma^{2}grams is seriously hampered by variations in the retention times along both axes. The peak positions may show small, random variations, as well as systematic drifts over a longer period of time. Shellie *et al.* concluded after an inter-laboratory study that current state-of-the-art GC×GC instruments are capable of achieving very impressive results

in terms of retention-time stability [84]. The instruments considered were equipped with cryogenic modulation, electronic pressure control and automatic injectors. However, operating such an instrument at or near the specified temperature limit of one of the analytical columns inevitably results in some degradation (stripping, ageing) of the stationary phase. In addition, retention times depend on component concentrations (non-linear isotherms) and residual material from the samples may also change the properties of the column. All these effects alter the behaviour of the GC×GC system and affect the observed retention times.

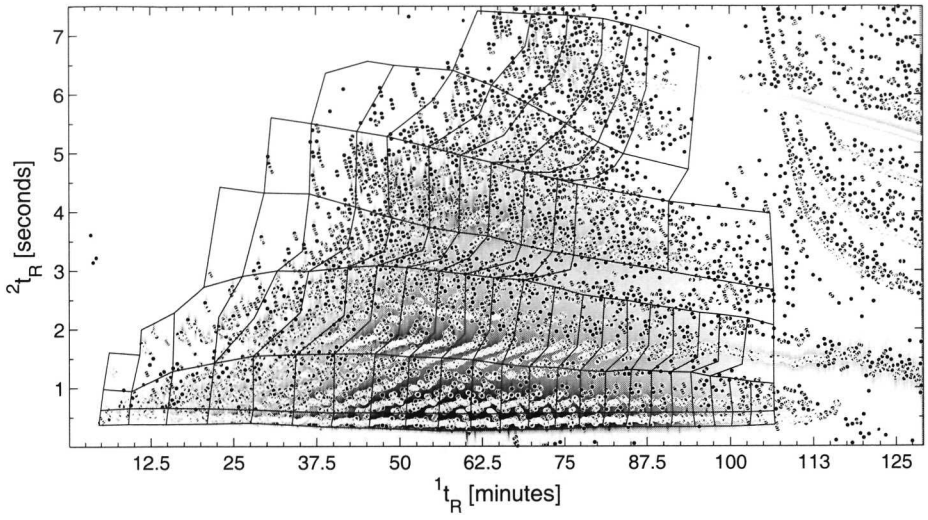


Figure 4.13: Two-dimensional chromatogram of the same diesel sample as in Figure 4.12 (recorded at 240 kPa inlet pressure).

Eliminating retention-time shifts is obviously very important. However, the best way to do so heavily depends on the objectives of the analysis. If sample profiling is the goal of the analyses, then the entire chromatographic 'Fingerprint' needs to be aligned (Chapter 5 of this thesis).
In Figure 4.12 and Figure 4.13, two chroma^{2}grams of a diesel sample are shown. The two samples were analyzed on the same instrument, using the same column combination and operating conditions. The only difference between the two chromatograms was the column inlet pressure: Figure 4.12 was recorded at 250 kPa, whereas for Figure 4.13 an inlet pressure of 240

kPa was used. This difference in pressure was intended to simulate the differences that may result from changing the column-set. The first chroma^{2}gram was quantified using an integration template. For this specific sample, the analytical method required 104 individual component groups to be distinguished based on the class of analytes and the number of carbon atoms.
Obviously, replacement of a column-set results in changes in the observed retention-times. Even small differences can render a carefully constructed integration template useless. Constructing a new integration template is a very laborious exercise. Alternatively, adjusting the chromatographic conditions in such a way that all components will elute at their original positions is a theoretical possibility. However, this is not currently feasible for GC×GC separations and the approach would likely be equally laborious to constructing a new template. Adaptation of the template has therefore been investigated as an elegant and effective alternative. An integration template consists of a number of integration polygons. Each individual polygon consists of a number of coordinates. In our approach, a virtual box is drawn around the integration template. The four corners of this box have coordinates representing maximum and minimum x (^1t_R) and y (^2t_R) values. The coordinates x and y are integer values, representing the position in the matrix. The four points are the lower-left corner $(min(x), min(y))$, upper-left corner $(min(x), max(y))$, lower-right corner $(max(x), min(y))$ and the upper-right corner $(max(x), max(y))$. Using these four locations, shifts and transformations can be calculated. The three transformations we have applied are shifting (of the complete template or a selection of the template), stretching or shrinking (in four directions), and shifting (in four directions) of each of the four corners of the box. All transformations should be performed such that integer values for both x and y result.
Shifting. Shifting of the template is straightforward. The points of the polygons all correspond to a set of x and y coordinates. The addition or subtraction of an integer to all these coordinates effectively shifts the entire template (all polygons) by a number of points in either direction. In this way, the template can be moved in four directions.
Stretching/shrinking. Stretching or shrinking of the template can also be performed in four different directions. In this approach, the box around the template is used to determine the minimum and maximum in the direction of the stretching. During stretching or shrinking, one of the lines determining

the box can be moved, while the three other lines remain in the same position. For the polygon coordinates between the two lines, linear interpolation is used to calculate their new position. The same approach is used when stretching is used and in other directions.

Corner stretching. Shifting and stretching or shrinking are not always sufficient to adapt templates to new chromatographic conditions. Modification of the four corners of the template can be a useful additional transformation. In this step, the four corners of the box around the template can be moved individually, while the other three corners are fixed. Again, linear interpolation is used to calculate the new positions of the points of the template. Since the box is transformed in two directions, linear interpolation must also be performed in two directions.

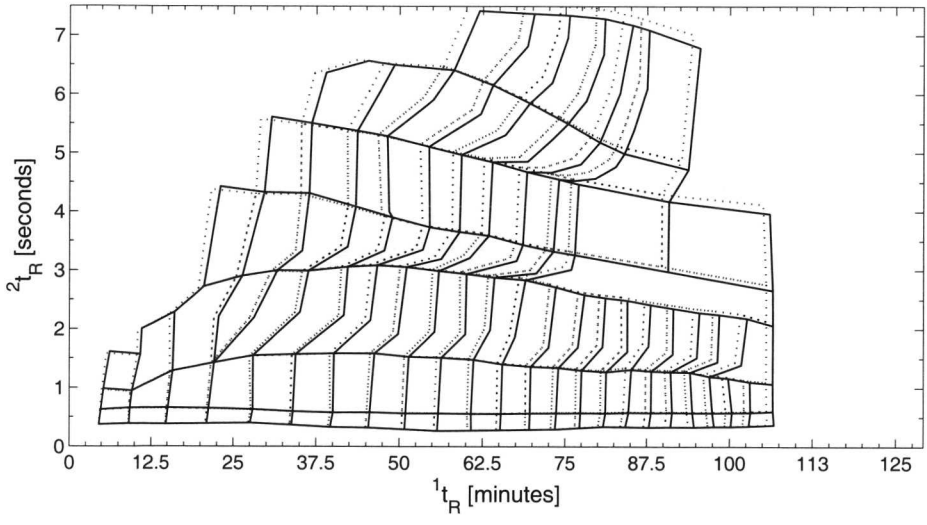


Figure 4.14: Effect of linear transformations on the integration template. The dashed line represents the original template as applied in Figure 4.12. The solid line indicates the template after transformation, which should be applied in conjunction with Figure 4.13.

Template modifications are carried out through visual inspection of the template superimposed on the chromatographic data. This is an interactive process, which allows instant evaluation of the results of actions performed

to adjust the template. Figure 4.14 shows the differences between the two templates. The solid lines represent the new integration template, created by modification of the original (dashed lines) integration template. The quantitative effect of the template modification is summarized in Table 4.1. For fifteen component groups the calculated concentrations of the various groups (in area%) before and after modification of the template is given.

Group nr.	Sample @250 kPa	Sample @240 kPa Template 250 kPa	% error[a]	Sample @240 kPa Mod. template	% error[a]
5	1.7	0.83	-53%	1.78	0.6%
6	2.93	3.85	31%	2.91	-0.7%
7	4.53	2.24	-51%	4.52	-0.2%
8	5.96	8.19	37%	5.90	-1.0%
9	6.01	3.22	-46%	6.17	2.7%
10	6.15	8.33	35%	6.14	-0.2%
14	1.90	1.13	-41%	1.98	4.2%
15	1.50	2.27	51%	1.42	-5.3%
23	0.07	0.09	29%	0.06	-14%
65	0.23	0.19	-17%	0.24	4.3%
75	0.38	0.21	-45%	0.33	-13%
81	0.25	0.35	40%	0.24	-4.0%
98	0.05	0.03	-40%	0.05	0%
102	0.27	0.20	-26%	0.27	0%

[a] Assuming that the results obtained by applying the original template on the original data are correct.

Table 4.1: The difference in the concentrations before and after modification of the template can be as much as a factor two. After modification of the templates very similar results are obtained from the two chroma^{2}grams.

4.4 Conclusions

Comprehensive two-dimensional gas chromatography has proven to be a very valuable separation technique for a large variety of applications. Because of the possibility to generate structured chromatograms, GC×GC is especially useful for the separation of component groups (group-type separations). However, a two-dimensional separation requires several generic steps for pre-processing of the data. Since the detector signal is a two-dimensional (intensity *vs.* time) signal, demodulation is required to obtain the three-dimensional surfaces that can be visualized as colour

or contour plots, the so-called chroma^{2}grams. Demodulation obviously also affects the locations of the peaks (*i.e.* the first- and second-dimension retention times). Another generic pre-processing step is baseline correction of the chromatograms. Baseline drift is typically manifested as an increase in the detector signal with increasing first-dimension retention times. The origin can be found in the linear time-intensity signal recorded by the detector. The proposed baseline-correction tool makes use of the linear detector output, as well as of the modulation sequence present in this signal. This tool clearly enhances the visibility of components present at low concentrations. The third generic tool is referred to as splining. In this step a set of user-defined peaks (typically a homologue series) is horizontally aligned at an arbitrary (usually low) second-dimension retention time. Splining corrects for variations in the column dead-time. During a temperature-programmed run with a constant inlet pressure the column dead-time increases significantly. Splining simplifies the interpretation of the chroma^{2}grams.

Group-type separations focus on component groups, which may contain large numbers of individual components. Quantification of such component groups, which is a very important aspect of group-type analysis, requires quantification tools different from those typically used for the quantification of individual (target) components. Our quantification procedures are based on conventional, reliable and readily available integration software. Demodulation of the signal results in retention coordinates rather than in retention times. Selecting a (polygonal) region for a component group is the first step in the quantification of the component groups. A straightforward summation of all peak areas in the selection box yields the relative area (proportional to the concentration) of that group.

Quantification of large numbers of component groups in very complex samples can be performed by constructing so-called integration templates. Such templates enable very fast quantification of similar samples measured under identical conditions. Changes in the retention times will, for example, always occur upon installation of a new column-set. Such changes render integration templates useless. A set of stretching and shifting routines allow adaptation of the template in such a way that it matches the data obtained under the new conditions. For two chromatograms recorded at inlet pressures of 250 and 240 kPa these routines were adequate for

matching the "old" template to the "new" chroma^{2}gram. Without changing the template, average quantification errors of 16% were observed, while the modified template showed average errors in the relative areas of only 5%. Modification of the quantification template is a strategy to eliminate retention-time shifts. Alternatively, retention-time shifts can be eliminated using the (second order) polynomials described in Chapter 5 of this thesis. However, these two approaches aim at different types of applications and for that reason cannot be compared.

Acknowledgements
The authors would like to acknowledge Nigel Wilson (ICI) for providing the alcohol samples. In addition, we would like to thank Jens Dallüge (Albemarle) for discussions on the template alignment.

Chapter 5

Alignment of GC×GC chromatograms.

The combination of multivariate analysis (MVA) and gas chromatography (GC) has been applied to a variety of applications. However, the success of this combination has been rather limited. By far the greatest impediment are retention-time shifts, which are inevitable in separation techniques. For conventional, one-dimensional GC several solutions have been proposed to eliminate, or at least drastically reduce, such shifts in retention time.
Comprehensive two-dimensional gas chromatography (GC×GC) offers a tremendous increase in peak capacity in comparison with conventional, one-dimensional GC. The resulting very detailed GC×GC chromatograms (or "chroma^{2}grams ") can be regarded as highly detailed fingerprints of a sample. This makes GC×GC a very attractive technique for the application of MVA. However, in a two-dimensional separation system retention-time shifts can (and will) occur in both separation dimensions. The successful combination of MVA with GC×GC therefore requires alignment techniques to eliminate retention-time shifts in both dimensions.
In this Chapter we will demonstrate the applicability of image-processing techniques for drastically reducing retention-time shifts for chroma^{2}grams. MVA techniques, such as PCA and Parafac, are used to quantitatively assess the results of the alignment. Parafac2 is demonstrated as an alternative method. In this case the retention-time shifts are corrected for within the algorithm. The three methods are successfully applied for reducing the

retention-time shifts present in two sets of chromatograms, one obtained by GC×GC with flame-ionization detection and the other from GC×GC with time-of-flight mass spectrometry. In addition, we demonstrate that the quantitative information is not affected by the proposed MVA methods.

5.1 Introduction

Gas chromatography (GC) is a very powerful tool for the quantitative and qualitative analysis of complex, volatile mixtures. In quantitative analysis, a number of relevant peaks are quantified, normally with the aid of integration software. The resulting quantitative information is, for example, required to meet legislation, for product specification or for waste monitoring. Qualitative analysis often involves the visual comparison of chromatograms, in which each chromatogram can be regarded as a chemical profile or fingerprint of a sample. Such a visual comparison is clearly very subjective. For a more-objective comparison MVA techniques can be applied. The systematic comparison of a large number of chemical profiles (*e.g.* gas chromatograms) with MVA techniques can yield valuable information on the differences or similarities between the samples. Eventually, this information can be linked to performance parameters [28] or it can be used for quality-control purposes [112].

Unfortunately, a straightforward application of MVA methods to chromatographic profiles is generally not possible. The greatest impediment is the retention-time instability associated with every analytical separation technique. For several reasons (see below) gas chromatograms exhibit small, but inevitable variations in the retention times. When applying MVA techniques constant (or "parallel") elution profiles are assumed, *i.e.* components are assumed to always elute at identical retention times with identical peak shapes. In practice, repeated analysis of a single sample will result in some variations in the retention time for any given component in the series of chromatograms. By using good, state-of-the-art instrumentation and sound (injection) procedures the degree of variation can be reduced substantially. The use of retention-time-locking algorithms can further improve these results [113]. However, because of the very nature of the chromatographic process variations in retention times and peak shapes can never be completely eliminated. Any variations in retention times and peak profiles will

be interpreted incorrectly by MVA techniques as changes in chemical composition.

There has been a great deal of interest in solving this problem, since it forms a major bottleneck for the application of MVA techniques to chromatographic data. One way to deal with retention-time shifts after the chromatograms have been recorded is by applying so-called data pre-processing techniques. This implies that chromatograms are corrected computationally before the data are subjected to MVA. Typically, the time axis of each chromatogram is altered in such a way that the result fits the chemical profile of a previously defined target chromatogram. Malmquist [37], Nielssen [38], Johnson [39], and Eilers [40] have described various techniques for the alignment of one-dimensional separation profiles. These techniques allow MVA to be applied on the corrected chromatographic data.

Three practical sources of shifts in retention times can be distinguished. Firstly, variations in operating conditions (*e.g.* flow or pressure, temperature) result in variations in retention times. Secondly, degradation of the stationary phase may occur. This can either be caused by gradual disappearance of the stationary phase ("phase stripping") or by chemical changes in the stationary phase by, for example, residual material in the sample. Thirdly, shifts will arise when replacing the column or by changing to another instrument ("method transfer") [113]. In addition, there are fundamental reasons why retention times and peak profiles vary in chromatography. Any non-linearity of the distribution isotherms will result in concentration-dependent times and profiles. Any influence of other analytes, matrix components, solvents, *etc.* on the distribution isotherms will also result in variations. We can (and should) try to approach ideal chromatography by creating excellent columns, avoiding secondary retention mechanisms (*e.g.* adsorptive surfaces), reducing the sample size, *etc.*. However, again, we can minimize the variations, but we cannot completely eliminate them. Generally, we wish to apply chromatographic analysis to diverse samples, with greatly varying concentrations and, possibly, composition. Trying to minimize concentration-dependent retention-time shifts by minimizing changes in the sample composition is defeating the purpose of chromatographic analysis.

The first practical source of shifts can largely be eliminated by using advanced instrumentation, such as auto-injectors and electronic pressure control. This reduces the variation in the injection time and offers a more stable

column flow, respectively. The second source of shifts, stripping of the stationary phase and chemical modification, can be reduced by using highly pure and effectively immobilized (cross-linked) stationary phases and by using pure (oxygen-free) carrier gases. Sample-induced chemical modification can be reduced by using suitable injectors (*e.g.* PTV) and liners. However, the threat cannot be completely eliminated. Unfortunately, since each component class responds differently to chemical modification of the stationary phase, the resulting shift is component-dependent. In extreme cases, the elution order may change. In addition, peak shapes can be altered. The third source of shifts (different columns) may be more easily overcome. Apart from avoiding the need to change the columns by using good procedures and materials (carrier gases and solvents), the effects of changing the column may often be corrected for. A new column with a slightly different diameter, stationary-phase thickness, and/or length results in a shift of all components in the same direction to different, but gradually varying extents. There are two routes towards solving this problem. The first option is to adapt the chromatographic conditions in such a way that components again elute in their original positions. Alignment of already recorded chromatograms is the second option.

In the last decade a novel separation technique has been introduced, *viz.* comprehensive two-dimensional gas chromatography (GC×GC) [1–3]. This technique offers a tremendously increased peak capacity in comparison with conventional, one-dimensional GC, because every part of the sample is subjected to two different separations. The value of GC×GC has already been demonstrated by a large variety of applications, such as oil and petrochemical products [64, 77, 107], halogenated compounds [70], fatty acids [68, 69], food analysis [62], cigarette smoke [114], essential oils [108] and environmental pollution [109]. The large peak capacity makes GC×GC a seemingly ideal technique in combination with MVA. The very detailed two-dimensional chromatograms (or chroma^{2}grams) can be regarded as highly detailed fingerprints of a sample.

Chroma^{2}grams have already been subjected to MVA techniques, for instance for the successful deconvolution of overlapping peaks [83], for enhancing detection limits [16], and for fast quantification [82, 115]. In all these cases retention-time shifts in the chroma^{2}grams were eliminated, or at least reduced, by applying shifts in local regions of the chromatograms in a data-

pre-processing step. The alignment procedures used in these studies can be regarded as local optimizations. In contrast, the elimination of shifts throughout the entire chromatogram, *i.e.* on a global scale, is much more difficult to achieve. As in conventional one-dimensional GC, shifts can partly be overcome by improved instrument electronics, such as pressure control and auto-injectors. In GC×GC state-of-the-art (cryogenic) modulators provide an excellent run-to-run repeatability [84]. For a sample containing 43 components (with concentrations varying from traces to high levels) six-replicate analyses were performed on a single column set. The authors reported an average retention-time repeatability of 0.12% (r.s.d.) in the first dimension and 0.74% in the second dimension, which are impressive results by GC standards. However, the use of a different column set (with nominally identical dimensions) led to significant shifts in the retention times in both dimensions [84]. Minute differences in column length, internal diameter, and stationary-phase thickness were suggested to have caused these shifts. The experimental run-to-run repeatability under perfect conditions on a single column set is difficult to improve by using alignment (pre-processing) techniques. The variations in the peak positions entail only one or two data points in either direction. Correcting for such minute differences on a global scale can easily result in over-compensation and in a deterioration of the retention stability. A multivariate model can be constructed based on the raw, unaligned chromatographic data. Any significant change in the conditions or the introduction of a new column-set, however, renders this model useless, since the chromatographic behaviour becomes different. A transfer method from one column-set to another would enhance the applicability of the model. Unfortunately, the global alignment of complete two-dimensional chromatograms has not yet been reported.

One way to overcome this problem is to develop models capable of handling chromatographic shifts. Bro *et al.* proposed the Parafac2 model for this purpose [100]. This model is only applicable to tri-linear data, such as a set of stacked chroma^{2}grams, and it is not applicable to conventional chromatograms. Instead of using elution profiles as such, the Parafac2 model uses a covariance matrix of the elution profiles. By doing so, the "inner-product structure" of the chromatograms is preserved. Parafac2 is not an alignment procedure, but it is an MVA technique with some tolerance for retention-time shifts.

In the field of image processing, transformation techniques are used to transform all kinds of images [116]. We have attempted to use such image-processing techniques on chroma^{2}grams, with the aim of global alignment. The results have been assessed with MVA techniques, such as PCA and Parafac. The effects of the alignment on quantitative data have also been examined. Finally, the image-processing techniques have been compared to the use of the Parafac2 method for dealing with retention-time shifts.

5.2 Theory

5.2.1 Comprehensive two-dimensional gas chromatography

In a two-dimensional-chromatography system, the effluent from the first dimension is passed through a modulation capillary. This device continuously traps and releases small portions of the effluent. In contemporary designs the modulator usually consists of two cooling jets along the modulator capillary or one jet, which is effectively used at two different locations in a loop design. The cooling gas is either evaporated nitrogen or expanding carbon-dioxide. Eluting components are trapped at the first cold spot. The trapped components are remobilized periodically by switching off or deflecting the cold jet. The pulsed portions are refocused by the second jet. Remobilization from the second jet constitutes the actual injection onto the second-dimension column. The detector at the end of the system registers the effluent from the second-dimension column. The detector output is one large string of second-dimension chromatograms.

Alignment procedures designed for one-dimensional chromatographic techniques may also be applied to two-dimensional separations. The signal obtained from a GC×GC instrument is essentially a time-intensity function, similar to a conventional, one-dimensional chromatogram. Identical features in the sample and target chromatograms can be used to create a synchronization profile. However, such an approach neglects the concealed chromatographic information of the modulated first dimension. Moreover, techniques that align the linear signal will not recognize the individually modulated peaks that belong to the same chemical component. Aligning the linear GC×GC signals is clearly not the most-appropriate approach. Alignment of the demodulated, two-dimensional chromatogram is preferred. Unfortu-

nately, GC×GC chromatograms can contain so-called wrap-around. Wrap-around occurs when the second-dimension retention time exceeds the modulation time (*i.e.* the duration of one modulation cycle) and it is reflected in spurious, broad peaks in subsequent second-dimension chromatograms. Using image-processing techniques, wrap-around and peaks eluting at or across the bottom and top edges of the chroma^{2}grams cannot be dealt with, since such techniques do not connect a point at the top of the chromatogram to a point at the bottom in the next column of data. However, alignment of the chroma^{2}gram does allow us to correct first- and second-dimension retention-time shifts simultaneously.

5.2.2 Image registration

Image-processing techniques are used in a wide variety of applications, such as image enhancement, image deblurring, image filtering, edge detection, and image transformation. Especially image-transformation techniques appear to be relevant in the present context. Such techniques are, for example, used in aerial photography. Aerial photographs are often registered from different perspectives (*i.e.* positions). For a correct comparison of the different images, the projection error must be eliminated. For this purpose, a *projective correction* method can be used [117].
Image transformation requires two images, referred to as *base* and *input*. The *base* or reference image is compared to the *input* or target image. The input image will be transformed, after which it is referred to as the aligned image. The first step is to register the two images. In this process, control points are selected in the two images. These are referred to as 'landmarks' and they are uniquely identifiable points in the two images. The coordinates of these control-points are used to calculate a transformation function between the two sets of points. The global transformation function used for this transformation is a mathematical expression, which is applied to transform the entire image. Obviously, the type of transformation function determines the flexibility, and the behaviour in case of extrapolation. For example, a higher-order polynomial can result in an excellent fit for the selected points, but can show strange anomalies in extrapolated regions. Transformation profiles applied to chromatographic data should allow non-linear corrections, but should exhibit a smooth behaviour. A polynomial function, therefore, seems

to be appropriate.
In a two-dimensional image, the shift is affected by the X-position, the Y-position, and possibly by the combined XY-position (correlation effect). In the form of a second-order polynomial, this yields Equations 5.1 and 5.2.

$$[X_{new}] = (a_x + b_x[X_{old}] + c_x[Y_{old}] + d_x[X_{old}Y_{old}] + e_x[X_{old}^2] + f_x[Y_{old}^2]) \quad (5.1)$$

and

$$[Y_{new}] = (a_y + b_y[X_{old}] + c_y[Y_{old}] + d_y[X_{old}Y_{old}] + e_y[X_{old}^2] + f_y[Y_{old}^2]) \quad (5.2)$$

Since the shifts are different in the X and Y directions, different values for the two sets of coefficients [a_x through f_x] and [a_y through f_y] are needed to describe the most-appropriate transformation profile.

5.2.3 Quantifying similarity of chroma^{2}grams

The effect of each alignment procedure may be characterized by a measure of similarity. In the case of two-dimensional data, a straightforward correlation coefficient clearly falls short. A two-dimensional analogue may be the 'inner-product correlation' [42].

$$r_{(\boldsymbol{A},\boldsymbol{B})} = \frac{tr(\boldsymbol{A}^T\boldsymbol{B})}{\sqrt{tr(\boldsymbol{A}^T\boldsymbol{A}) \times tr(\boldsymbol{B}^T\boldsymbol{B})}} \quad (5.3)$$

Where:

$r_{(A,B)}$	Correlation coefficient between matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$
$\boldsymbol{A}$	Standard matrix
$\boldsymbol{B}$	Sample matrix
tr	Trace function (sum of all diagonal elements)

This measure has already been applied successfully for quantifying the effect of shifting within local regions in chroma^{2}grams (Chapter 3 of this thesis) and it should also be applicable to entire chroma^{2}grams. However, considerably more computational effort will be required. For the comparison of a large set of chromatograms, the approach described above is not attractive, since it results in a matrix of correlations for every chromatogram relative to each of the other chromatograms. Multivariate-analysis techniques are perfectly suited for comparing large numbers of

objects (chroma^{2}grams in this case).

PCA

The results of alignment procedures can be quantitatively evaluated using MVA routines. For one-dimensional chromatography, principal components analysis (PCA) can be employed. In PCA, the original variables are replaced by a (strongly) reduced number of uncorrelated (orthogonal) variables, called the principal components. Mathematically:

$$\boldsymbol{X} = T \times P^T + \boldsymbol{E} \tag{5.4}$$

Where:

$\boldsymbol{X}$	Original dataset containing n (samples)$\times$ p (variables)
T	scores of n (samples) $\times$ F (principal components)
P^T	transposed loadings containing F (principal components) $\times$ p (variables)
$\boldsymbol{E}$	Residuals, variation not explained by the model

The principal components are constructed in such a way, that the first one (PC1) represents the main source of variation in the original dataset. The second PC is orthogonal to the first one and it represents the maximum variance not explained by PC1. Each PC is a linear combination of the original variables. The direction of each PC in the original variable space is expressed in the principal-component loadings.
The number of PC's provides an indication of the complexity of the model. If the data are highly correlated, a few PC's will be sufficient to reproduce the original data. A way of presenting the data obtained by PCA is the score plot. Related objects (belonging to the same group) have similar scores on the PC's and will consequently tend to cluster. Since the alignment of chromatograms should be evaluated on identical samples, there is no source of variance from the sample. The only source of variation between two (sets of) chromatograms is their chromatographic behaviour. Ideally, the comparison of two sets of chromatograms, measured with two columns, would result in a PCA model in which most of the (mathematical) variance is captured in one or two principal components. After alignment, the main source of variance between the two sets is captured in the first principal component.
A limitation is that PCA can only deal with matrices, *i.e.* sets of "one-

dimensional" chromatograms. Data matrices as encountered with GC×GC have to be "remodulated" or linearized. The resulting "one-dimensional" chromatograms can then be subjected to PCA. From this perspective, multiway methods, such as parallel factor analysis form an obvious alternative. These methods can deal with sets of data matrices, instead of data vectors.

Parafac

Parallel factor analysis (Parafac) is a generalization of PCA towards higher orders. It is a true multiway technique, which decomposes a multiway dataset into one or more combinations of vectors ("triads"). The Parafac model was proposed in the 1970's independently by Carrol and Chang under the name CANDECOMP (canonical decomposition) [97] and by Harshman under the name Parafac [98]. Essentially, Parafac models the data as follows:

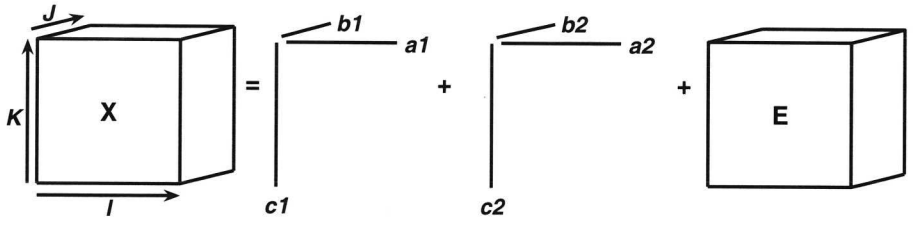


Figure 5.1: Schematic two factor Parafac model.

In this schematic overview, the stacked chromatograms are represented by the matrix $\boldsymbol{X}$ with dimensions $(I \times J \times K)$. In our case I indicates the first-dimension retention time, J the second-dimension retention time, and K the specific sample or injection. Analogously to PCA, the effect of the alignment procedure may be evaluated from the percentage of variance captured in the first parallel factor.

Parafac2

Most multiway methods assume parallel proportional profiles (*e.g.* invariable absorption wavelengths or elution times). In some cases, unequal record lengths may need to be dealt with, such as in batch-process analysis, where the time required to process a batch may vary, resulting in unequal record lengths. In chromatography, peaks may shift due to practical or

fundamental reasons. Most multiway methods cannot deal with such shifts. Parafac2 can handle shifted profiles through the inner-product structure. This property can be used, for example, to deal with stretched time axes [101]. The Parafac2 algorithm can be described schematically as follows:

$$\boldsymbol{X}_k = \boldsymbol{A}_k \boldsymbol{D}_k \boldsymbol{B}^T + \boldsymbol{E}_k \tag{5.5}$$

Where:

X_k	Chroma^{2}gram of the k^{th} sample ($I \times J$)
$\boldsymbol{A}_k$	Matrix containing ${}^1t_R$ elution profile the for k^{th} sample ($I \times R$)
$\boldsymbol{D}_k$	Diagonal containing weights (relative concentrations) of k^{th} sample of $\boldsymbol{X}$ ($R \times R$)
$\boldsymbol{B}$	Matrix containing ${}^2t_R$ elution profiles ($R \times J$)
$\boldsymbol{E}_k$	Residual for k^{th} sample in $\boldsymbol{X}$ ($I \times J$)

A useful property of $\boldsymbol{A}_k$ is that $\boldsymbol{A}_k^T \boldsymbol{A}_k = \boldsymbol{A}^T \boldsymbol{A}$ for $k = 1, .., K$. In other words, the cross-product of the $\boldsymbol{A}$ matrix is constant for all samples. In literature, Parafac2 has been used for the decomposition of data obtained by liquid chromatography with photo-diode array detection [100] and for fault detection in batch-process monitoring [31]. Parafac2 only allows the inner-structure relationship to be used in one direction. For LC-PDA this is not a serious limitation, as retention-time shifts only occur in the LC direction. For GC×GC, however, shifts can (and will) occur in both directions and they are not identical along the two retention axes. In applying Parafac2 to chroma^{2}grams, the inner-structure relationship is applied along the first-dimension axis. In this direction, differences in peak shape are characterized by so-called "in-phase" and "out-of-phase" (*i.e.* the top of the first-dimension peak falls almost exactly in between two second-dimension fractions) between different injections [118].

5.3 Experimental

5.3.1 GC×GC-FID

Experiments of the GC×GC with an FID were performed with an Agilent 6890 GC (Wilmington, DE, USA). This GC was equipped with a CIS 4

programmed-temperature-vaporization (PTV) injector (Gerstel, Mulheim an der Rühr, Germany) and a CTC CombiPal (CTC Analytics, Zwingen, Switzerland) auto-injector. The modulator was a KT 2003 thermal modulator (Zoex, Lincoln, NE, USA) and the system was equipped with a separate second-dimension oven, which allowed flexible temperature programming of the second-dimension column. Liquid nitrogen was used as the source for cold modulator gas at a flow of approximately 117 mL/min. The modulation time was 7.5 s and the duration of the hot pulse was 300 ms.
The temperature of the first-dimension column oven was programmed from 40°C (5 minutes isothermal) at a rate of 2.5°C/min to 250°C (20 minutes isothermal). The hot pulse of the release jet was set at 100°C above the oven temperature, while the second-dimension oven was operated at an offset of 50°C above the temperature of the primary ("first-dimension") oven.
The PTV injector was programmed from 40°C to 250°C (5 minutes isothermal) with a ramp of 12°C/s.
The column-set consisted of a of a 10 m length × 0.25 mm internal diameter 0.25 μm film thickness DB-1 column (J&W Scientific, Folsom, CA, USA) in the first dimension and a 2 m length × 0.1 mm internal diameter 0.1 μm film thickness BPX50 column (SGE, Ringwood, Australia) in the second dimension. A fused-silica capillary of 0.5 m × 0.1 mm deactivated with diphenyltetramethyl-disilazane (DPTMDS), obtained from BGB Analytik (Anwil, Switzerland) was used to connect the second-dimension column to the flame-ionization detector (FID). Columns were coupled with custom-made press-fits (Techrom, Purmerend, The Netherlands). In all experiments, helium was used as carrier gas.

Conditions set 1
Modulation was performed using a 1.6 m × 0.1 mm DPTMDS-deactivated fused-silica capillary (BGB Analytik). The inlet pressure was 250 kPa, resulting in a carrier-gas flow of approximately 1 mL/min at 40°C at the column outlet.
Conditions set 2
Modulation was performed using a 2.0 m × 0.1 mm DPTMDS-deactivated fused-silica capillary (BGB Analytik). The inlet pressure was 280 kPa, resulting in a column flow of approximately 1 mL/min at 40°C.

Sample

The sample used in this study was a synthetic mixture, containing all the components of the "Grob mix" [119]. This mixture includes C_9, C_{12}, and C_{15} linear alkanes, 2,3-butanediol, 2,6-dimethylphenol, 2,6-dimethylaniline, 2-ethylhexanoic acid, 1-octanol, dicyclohexylamine and methyldecanoate. To this mixture, toluene, decaline (both *cis* and *trans*), 2-methylnaphthalene and C_{16}, C_{17}, C_{19}, and C_{20} alkanes were added to have components eluting at longer first-dimension (alkanes) and second-dimension (aromatics) retention times. The concentrations of all components were approximately 500 ppm (weight). Cyclohexane (*p.a.* quality, Merck) was used as solvent. The injection volume was 1 μL, with a split flow of approximately 100 mL.

Instrument control and data processing

Instrument control and data acquisition were performed with EZ-Chrom Elite (v2.61, SSI, Willemstad, The Netherlands). Data were collected at 100 Hz to obtain a sufficient number of data points across a peak. Chromatograms were exported to the Common Data Format (CDF). Data handling was performed in MATLAB R14, service pack 1, including the Image Processing toolbox, version 5.0.1 (The Mathworks, Natick, MA, USA). Data-handling routines were developed in-house. In addition, the NetCDF toolbox [104]. Prior to further processing, the chroma^{2}grams were splined according to the procedure described in Section 4.2.3, on page 61 of this thesis.

5.3.2 GC×GC-ToF-MS

Experiments were performed on a Pegasus 4D system (ATAS, Cambridge, U.K.).

The column-set consisted of a of a 15 m length × 0.25 mm internal diameter 0.25 μm film thickness DB-5MS column (J&W) in the first dimension and a 1.2 m length × 0.1 mm internal diameter × 0.1 μm film thickness BPX50 column (SGE) in the second dimension. The modulation time was 4 s and the hot-pulse duration was 1600 ms.

The temperature for the primary (first-dimension-column) oven was pro-

grammed from 70°C (3 minutes isothermal) at a rate of 5°C/min to 300°C (5 minutes isothermal). The hot pulse of the release jet was operated at a temperature of 100°C above the oven temperature, while the secondary oven was operated at an offset of 30^oC above the temperature of the primary oven.
Injection was performed using an Atas Optic 3 injector in the hot-split mode at 260°C. The transfer line to the MS was kept at 325°C. The solvent delay was 300 seconds and the detector-scan range was 45 to 450 m/z. The detector voltage was 1750 V, the filament bias voltage was -70 V, and the ion source was kept at 280°C.
The four experiments were performed under a constant-flow regime. The flow settings were 1.0, 1.1, 1.2, and 1.3 mL/min (at 40 °C). However, the actual flow was difficult to determine. Helium was used as carrier gas. The sample was an oximated and sylilated plant extract (a chloroform extract of 10 g honeyfried glyccerhizae, 5 g ehedra, 5 g coicis and 4 g armeniacea).

5.4 Results and discussion

5.4.1 Repeatability

The first set of chromatograms, measured under the conditions specified for set 1, consisted of 24 injections of the test sample over a period of three weeks. In order to determine the repeatability, the retention coordinates in both dimensions of all of the 19 components in the sample were calculated. The resulting repeatability was comparable with the results reported by Shellie *et al.* [24]; the average relative standard deviation of D1 was found to be 0.06% (or 0.17 minutes) *, while for D2, the average r.s.d. was 0.84% (or 0.01 seconds). The similarity of the initial set of chromatograms was calculated using the inner-product correlation. The first 50 columns in the data matrix (6.25 minutes in D1) were discarded, since this region contains the solvent peak. Furthermore, the baseline was corrected by an algorithm described in Section 4.2.2 on page 59. The average inner-product correlation for these 24 chromatograms was found to be 0.933, indicating a high similarity. In Figures 5.2 and 5.3 an overlay of five chroma^{2}grams from this series

*The difference between two successive points in D1 is 7.5 seconds (t_M). Small differences in an "out-of-phase" peak may results in a difference in ^2t_R of 7.5 seconds.

is shown.

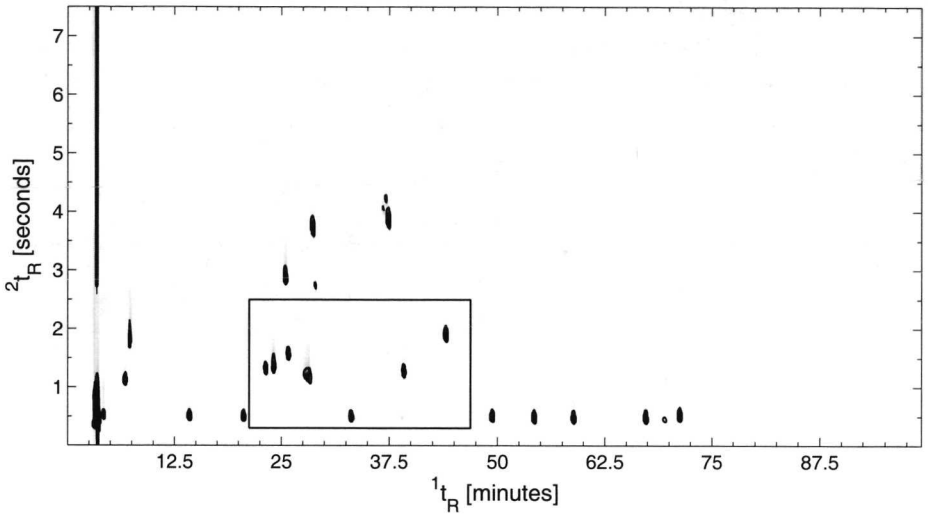


Figure 5.2: Overlay of five chroma^{2}grams acquired under the conditions of set 1.

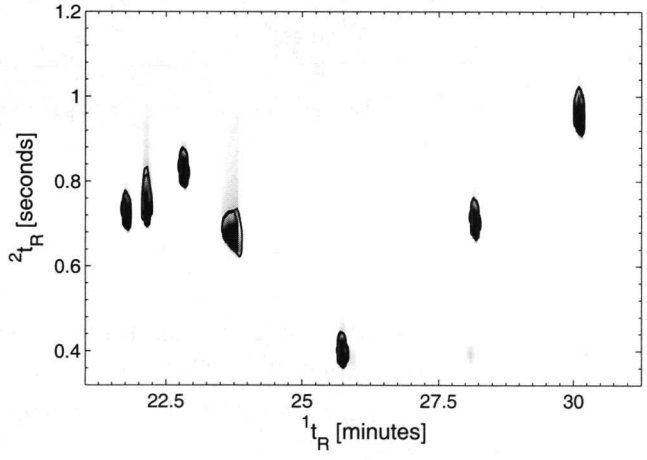


Figure 5.3: Enlargement of a region from Figure 5.2

The chroma^{2}gram from the first injection is shown in the form of a so-called colour plot (in grey-scales). For the other chroma^{2}grams, single contour lines

(at a certain peak height) were calculated and plotted on top of the initial chromatogram. The contours of the overlaid chroma^{2}grams closely match the peak shapes of the original chroma^{2}gram. This indicates a high retention stability. The second series of measurements on the same sample was measured at the conditions specified for set 2. It consisted of five consecutive injections, measured across two days. The average inner-product correlation of this set was somewhat lower: 0.795.

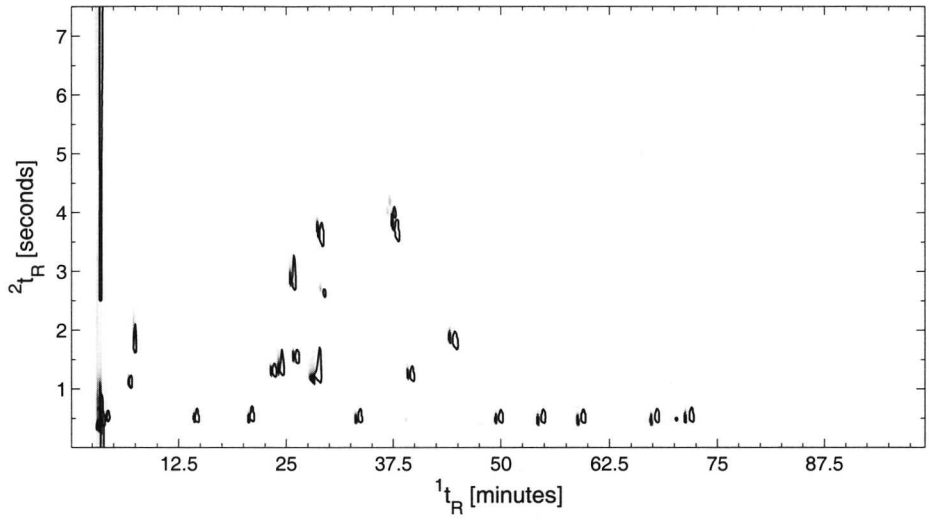


Figure 5.4: Peak contours obtained using column-set 2 plotted on top of a chroma^{2}gram obtained using column-set 1 for the same sample.

The differences between sets 1 and set 2 are reflected in Figure 5.4. Across the entire chroma^{2}gram there is a difference in both the first- and second-dimension retention times. The r.s.d. for retention times in D1 for the combined set was 0.7%, while in D2 it was 2.1%. The latter is mainly caused by the relatively large variation in the second set of five chromatograms. The reason for this greater variation is yet unclear.

5.4.2 Transformation profile

Using the image-registration tools from MATLAB, a set of eleven control points were selected for matching peaks in the two chromatograms. From

the eleven control points, six belonged to components in the sample, while the other five originated from contaminants present in all chroma^{2}grams. The actual shift between the control points in the two chroma^{2}grams is presented in a "velocity plot", shown in Figure 5.5 .

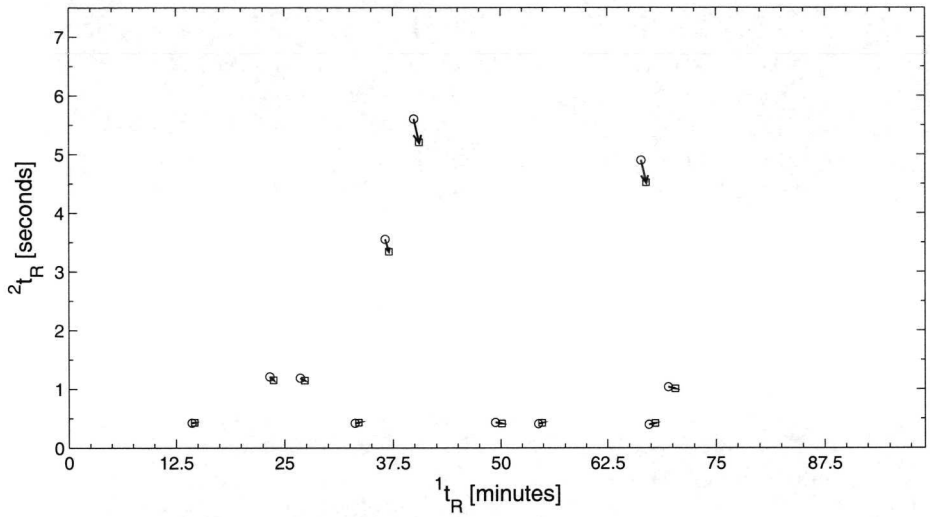


Figure 5.5: "Velocity plot", visualizing shifts in retention times between two chroma^{2}grams.

The two sets of eleven points represent the location of the control points in the *base* or reference chroma^{2}gram (squares) and in the *input* or target chroma^{2}gram (circles). The arrows indicate the direction and magnitude of the shift. This set of data was used to estimate the coefficients a trough f in the second-order polynomial Equations 5.1 and 5.2. From these control points, a transformation profile is derived using the MATLAB image processing toolbox. This yields the coefficients in both the X and Y directions. Since each equation requires six coefficients, the resulting matrix that represents the polynomial functions has the dimensions 6×2. For a given location $[X_{old}, Y_{old}]$ this transformation function will produce $[X_{new}, Y_{new}]$. The global transformation profile can be visualized by calculating the magnitude of shifts for each individual point in the data matrix and by projecting these shifts in a (grey-scale) colour plot. Figure 5.6 shows such a visualization. Unfortunately, the direction of the shift

cannot be presented in such a picture.

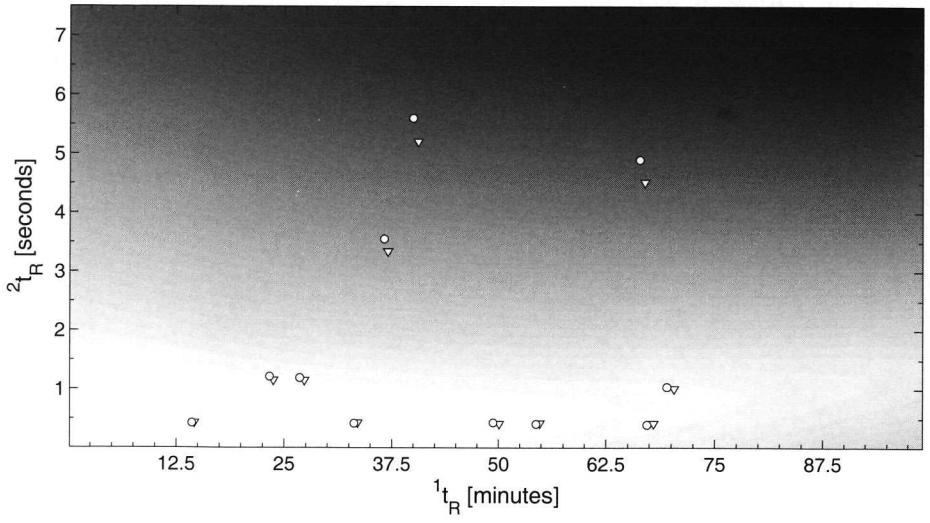


Figure 5.6: Transformation profile, showing the magnitude of the shifts between the two chroma^{2}grams.

From Figure 5.6 it can be concluded that the second-order polynomial performs a gradual shift in both dimensions.

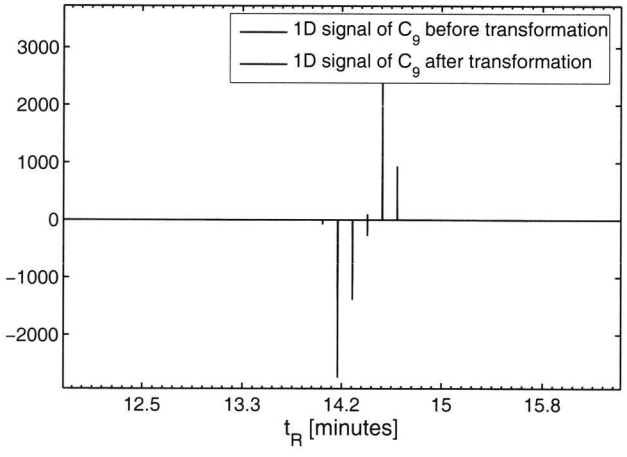


Figure 5.7: Effects of the transformation profile on the nonane peak, displayed in the form of the original modulated signal.

Towards higher retention times in D1 and D2 the magnitude of the shift increases.

The effect of the transformation on the recorded chromatographic signal is illustrated in Figure 5.7. The nonane peak (C_9) in the upper part of the chromatogram exhibits so-called in-phase behaviour. After transformation (lower part of Figure 5.7) in the two-dimensional domain, the peak position is shifted toward lower ${}^1t_R$, while the modulation sequence is altered such that it shows almost perfect out-of-phase behaviour.

Applying the transformation profile

The effect of the transformation profile was tested on representative chromatograms from both sets of data. The inner-product correlation of the chromatograms before and after transformation was used to select an optimal set of control points. Prior to image transformation, the inner-product correlation was less than 0.01, indicating that the two chroma^{2}grams were totally dissimilar.

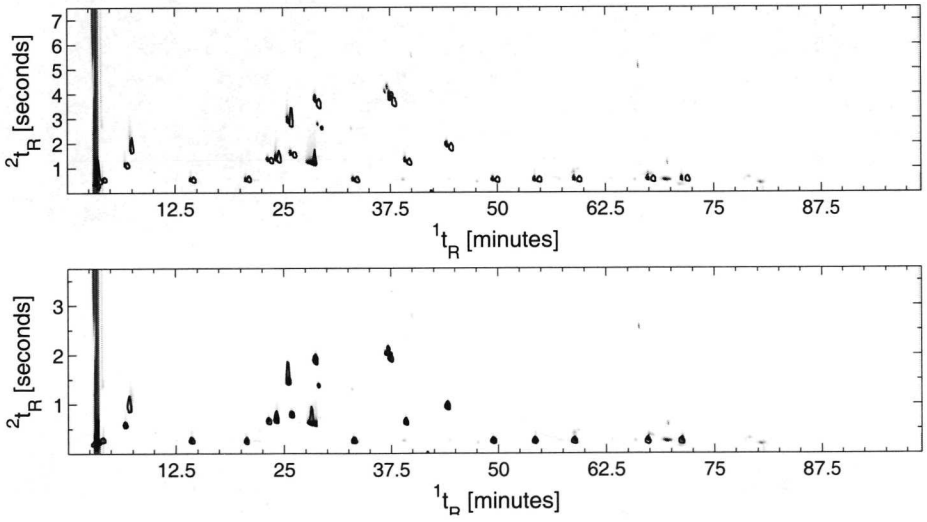


Figure 5.8: Effect of image transformation. The upper chroma^{2}gram shows the overlay of set 1 and set 2 prior to the transformation function. The lower chroma^{2}gram shows the overlay after applying the transformation profile.

After image transformation, the inner-product correlation was found to be

0.811. Overlays of the two chroma^{2}grams before and after alignment can be found in Figure 5.8. The main objective of the alignment procedure was to match the latter five chroma^{2}grams (new column set) to the former 24 chroma^{2}grams (old column set). The average inner-product correlation of the five "new" chroma^{2}gram with the 24 "old" chroma^{2}gram prior to the transformation was 0.059. After transformation of the data, the average inner-product correlation had been increased to 0.74.

5.4.3 Retention-time stability

The use of (higher order) polynomials for the transformation of retention times can result in anomalies in extrapolated regions of the data. The parameters of the second-order polynomials used above were estimated from a set of control points located in the middle of the chroma^{2}gram. To

Component	Reference	s.d. in ^1t_R before transform.[a]	s.d. in ^1t_R after transform.[a]	s.d. in ^2t_R before transform.[a]	s.d. in ^2t_R after transform.[a]
Nonane (C_9)	*M*	1.08	0.39	2.66	2.91
Decane(C_{10})	*I*	1.09	<0.01	0.93	0.97
Dodecane (C_{12})	*M*	1.45	0.13	0.80	0.93
Pentadecane (C_{15})	*M*	1.76	0.42	1.08	1.29
Hexadecane (C_{16})	*M*	1.81	0.01	1.50	1.75
Heptadecane (C_{17})	*I*	1.81	0.11	1.22	1.48
Nonadecane (C_{19})	*M*	1.84	0.38	1.26	1.49
Eicosane (C_{20})	*E*	2.17	0.10	1.34	1.54
2,3-Butanediol	*E*	0.72	0.37	3.41	1.74
1-Octanol	*I*	1.09	0.14	1.15	0.70
Toluene	*E*	0.72	0.26	3.91	3.25
cis Decalin	*E*	1.22	0.27	3.17	1.98
trans Decalin	*M*	1.18	0.33	2.47	1.73
1-Methylnaphthalene	*I*	1.81	0.29	8.29	3.09
2,6-Dimethylaniline	*I*	1.44	0.27	5.58	2.71
2,6-Dimethylphenol	*I*	1.19	0.26	3.95	1.58
2-Ethylhexanoic acid	*I*	2.54	1.41	1.34	1.11
Methyldecanoate	*I*	1.48	0.19	1.87	1.26
Dicyclohexylamine	*I*	2.09	0.55	5.51	3.65

[a] Standard deviation calculated over 29 peak positions, expressed in datapoints (recorded at 100 Hz).

Table 5.1: Standard deviations in peak apex coordinates before and after image transformation.

investigate possible extrapolation effects, the standard deviations of the

peak coordinates are used. In Table 5.1, the standard deviations for both ${}^1t_R$ and ${}^2t_R$ before and after the transformation are presented for all 19 components in the sample. Annotations indicate whether a peak was used as a marker ($\boldsymbol{M}$) to construct the transformation profile and should, therefore, be properly aligned, whether the new peak position was calculated using interpolation ($\boldsymbol{I}$), or whether the peak position was extrapolated ($\boldsymbol{E}$).
An improvement in the standard deviations of the peak coordinates is observed for almost all components. The improvement is generally much greater for ${}^1t_R$ than for ${}^2t_R$. This is due to the splining procedure, which acts like a synchronization step in the second dimension direction. For the n-alkanes the observed precision in ${}^2t_R$ is slightly worse after the transformation of the second set. This can be explained by the already rather small differences before the transformation. The overall conclusions are that the standard deviations in the peak positions (both ${}^1t_R$ and ${}^2t_R$) are improved and that extrapolation is not significantly worse than interpolation.

Area preservation
Alignment procedures should not affect the chromatographic information contained in the data.
However, (non-linear) transformation profiles aimed at changing peak positions will also result in changes in the peak shapes in D1. The peak area, which represents the quantitative information in the chromatogram, should not change in this process. To investigate the effect of the transformation, the relative peak areas for the 19 components were compared before and after transformation of the data. The results are presented in Table 5.2. The reference peaks (control points) are marked M, the interpolated peaks (located between the control points) are marked I and the extrapolated peaks (located outside the range spanned by the control points) are marked E. These results indicate that the transformation does not significantly affect the quantification of the 19 target components.

5.4.4 Effect of image transformation on MVA

To evaluate the effect of image transformation on the subsequent application of multivariate-analysis techniques, the complete dataset was subjected to PCA, Parafac and Parafac2.

Component	Reference	Area before transformation [area%]	Area after transformation [area%]
Nonane (C_9)	*M*	6.01	5.91
Decane(C_{10})	*I*	7.12	7.02
Dodecane (C_{12})	*M*	6.84	6.77
Pentadecane (C_{15})	*M*	6.46	6.45
Hexadecane (C_{16})	*M*	8.75	8.78
Heptadecane (C_{17})	*I*	6.87	6.88
Nonadecane (C_{19})	*M*	6.61	6.67
Eicosane (C_{20})	*E*	5.01	4.99
2,3-Butanediol	*E*	1.26	1.27
1-Octanol	*I*	5.00	4.99
Toluene	*E*	4.11	4.06
cis Decalin	*E*	2.59	2.57
trans Decalin	*M*	2.37	2.36
1-Methylnaphthalene	*I*	6.23	6.34
2,6-Dimethylaniline	*I*	6.18	6.27
2,6-Dimethylphenol	*I*	5.06	5.06
2-Ethylhexanoic acid	*I*	0.86	0.90
Methyldecanoate	*I*	4.61	4.60
Dicyclohexylamine	*I*	5.84	5.89

Table 5.2: Relative peak areas before and after transformation.

PCA

Prior to PCA chroma^{2}grams were "remodulated" (or unfolded) to yield a string of fast second-dimension chromatograms.

Figure 5.9 displays the results of PCA. The captured variances in PC1 and PC2 were 48% and 40%, respectively, before image transformation. After transformation, the percentage of variance contained in PC1 was 80%, while PC2 captured 11% of the variance in the data. Projection of the principal component scores also yielded the expected results. In the initial situation, prior to image-transformation, scores are projected exclusively on one of the principal-components axis. The axis of PC1 contains the chroma^{2}grams from the first set of experiments. On the second axis, chroma^{2}grams 25 to 29 are located. The position of the chroma^{2}grams on the two axes indicates no relation at all between the two principal components. This can be explained by the overlay in Figure 5.2. Peaks are shifted to such an extent that there is no overlap. This means that in regions in which peaks are found in set 1, no signal (and thus variance) is found in set 2.

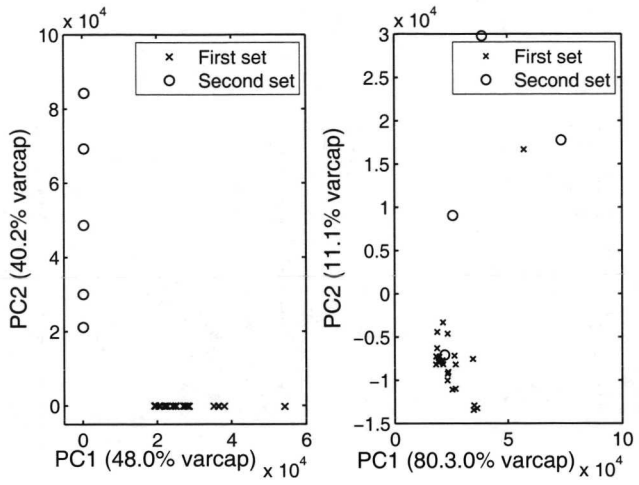


Figure 5.9: Results obtained by PCA before and after transformation. The score plot on the left side (*a*) shows the clustering prior to the transformation of the data. The score plot on the right side (*b*) displays the clustering after application of the transformation profile

The principal components of set 1 show no score in set 2 and vice versa. After transformation there is no distinct difference between the two groups. chroma^{2}grams 25 to 29 are still somewhat separated from the large cluster of other chroma^{2}grams. However, this is probably caused by differences in the peak shapes for the more-polar components in the mixture. Especially hexanoic-acid and 2,6-dimethylphenol exhibit different peak shapes in the second dimension on the two column-sets. Such differences may be caused by adsorption effects within or outside the columns. Alignment procedures cannot deal with such variations.

Parafac

Describing the chromatographic data with a two-component parallel-factor model resulted in similar score plots as obtained by PCA. Compared to Figure 5.9*b*, Figure 5.10*b* shows slightly more overlap between the two sets. However, the percentage variance described by the model was quite low. A one-factor Parafac model captured 32.0% of the variance before image transformation. For a two-factor model this increased to 60.2%. After the transformation step 56.8% of the variance was captured (63.7% in a

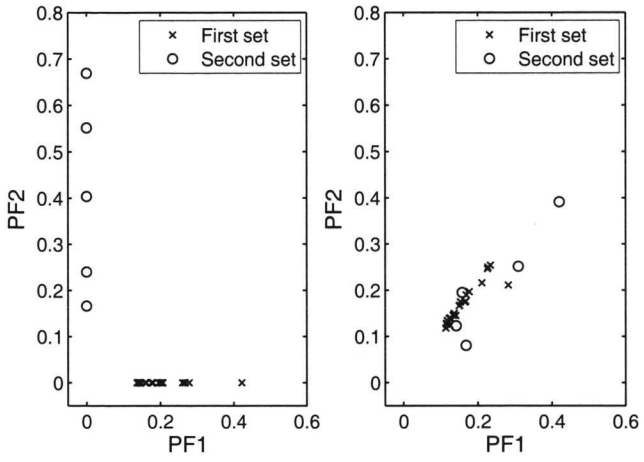


Figure 5.10: Results obtained by Parafac before and after transformation. The score plot on the left side (*a*) shows the clustering prior to the transformation of the data. The score plot on the right side (*b*) displays the clustering after application of the transformation profile

two-factor model).

Parafac2

Application of the Parafac2 model should, ideally, not result in different score-plots for the datasets before and after image transformation. The inner-product structure in both cases should be identical. For the dataset containing 29 chromatograms, the computational effort for the Parafac2 model is severe.

As can be seen in Figure 5.11, there is hardly any difference between the score plots before and after image transformation. Furthermore, both results are similar to the Parafac results after image transformation. This implies that the Parafac2 model is capable of dealing with retention-time shifts in both dimensions. Before and after image transformation, the captured variance is 69.3% and 68.7%, respectively. This is somewhat higher than the result obtained with the two-factor Parafac model after image transformation. There may be a major drawback in the use of the Parafac2 algorithm. The use of alignment techniques enables inspection of the direction and magnitude of the shift (Figures 5.6 and 5.7). Parafac2

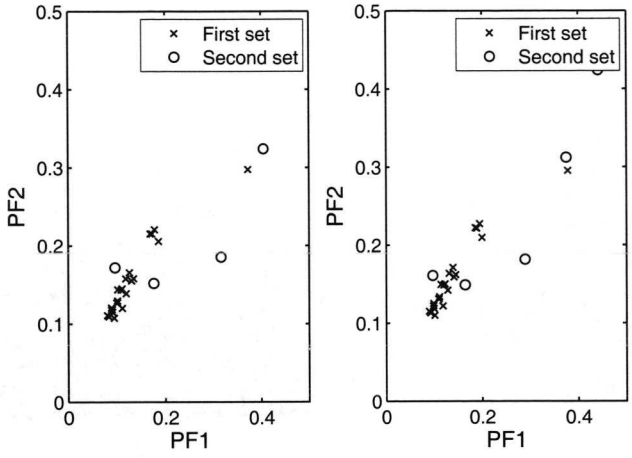


Figure 5.11: Results obtained by Parafac2 before and after transformation. The score plot on the left side (*a*) shows the clustering prior to the transformation of the data. The score plot on the right side (*b*) displays the clustering after application of the transformation profile

does apparently handle these shifts correctly, based on the results in Figure 5.11. It is however not clear if these shifts are dealt with exactly. The model therefore behaves like a "black-box". A second drawback is the computational effort required. Whereas the Parafac model took about one minute to converge, the Parafac2 model took about one hour. Furthermore, since the algorithm uses the covariance matrix, inspection of the factor loadings is not possible. This last drawback is significant, since it implies that no explanation can be given for the resulting scores.

5.4.5 GC×GC-ToFMS

A set of four chroma^{2}grams were recorded using a GC×GC-TOF-MS instrument at different, but constant flow rates. For this study, the total ion current (tic) was extracted from the CDF data files. Prior to further processing, these data files were subjected to baseline correction to eliminate baseline drift.

An overlay of the four chroma^{2}grams is presented in Figure 5.12. The first sample was used as a reference chroma^{2}gram. For each of the other three

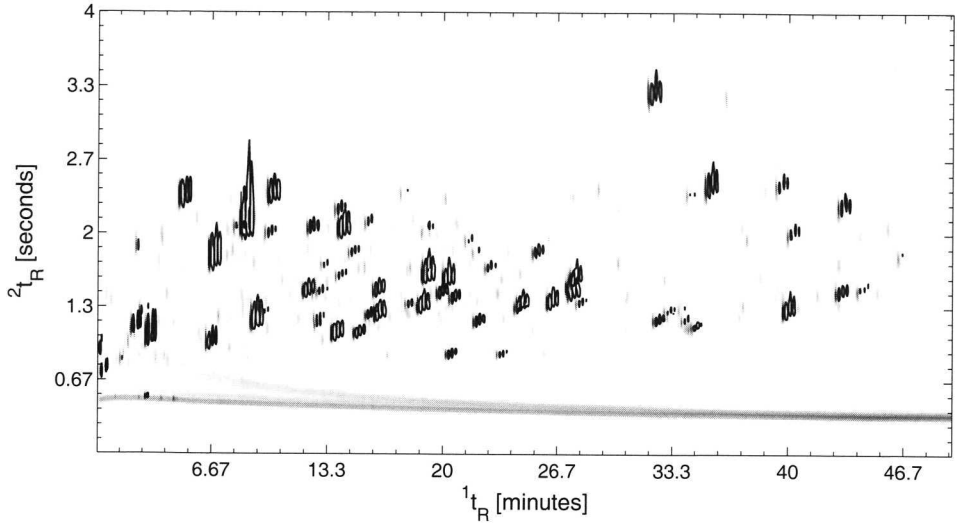


Figure 5.12: Overlay of four GC×GC-TOF-MS chroma^{2}grams (total ion current) before alignment.

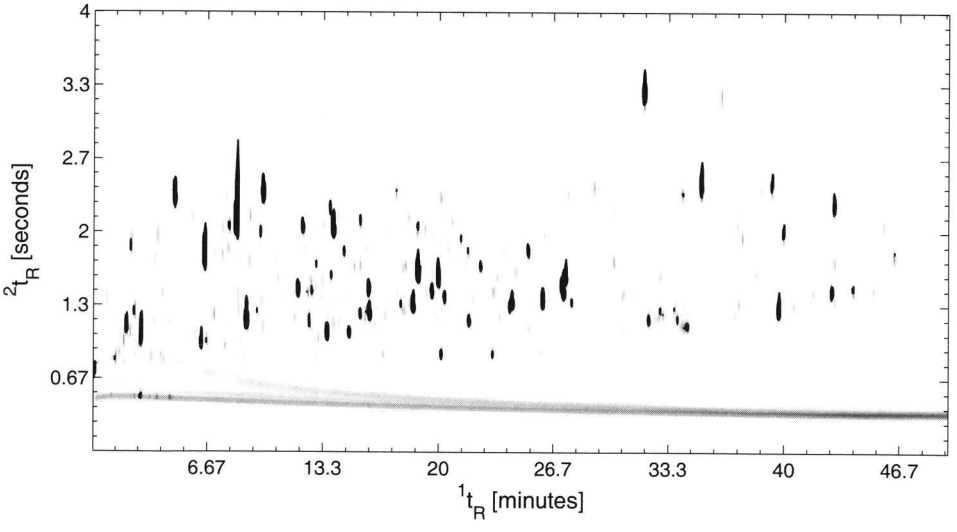


Figure 5.13: Overlay of four GC×GC-TOF-MS chroma^{2}grams (total ion current) after alignment.

chroma^{2}grams, transformation functions were calculated to match the first sample. Three transformation functions were constructed for this purpose, based on 18 reference points.
The average inner-product correlation between the four chroma^{2}gram before transformation was 0.11. After applying the image-transformation steps, the mean inner-product correlation increased to 0.77. This is also reflected in the overlay presented in Figure 5.13. When subjected to PCA, the captured variance in PC1 for the original data was found to be 58%. After image transformation, this was increased to 93%, all in the first principal component.

5.5 Conclusions

Variations in retention times in GC can be minimized by using state-of-the-art instruments and carefully controlled procedures. Method-transfer tools, such as "retention-time locking" can be used to further minimize the variations in conventional, one-dimensional GC. However, variations in retention times can never be completely eliminated and method-transfer tools do not yet exist for comprehensive two-dimensional gas chromatography (GC×GC). For the latter kind of data, image-processing techniques provide alignment tools in the form of image-registration techniques. This was successfully demonstrated for two sets of chromatograms obtained by GC×GC ("chroma^{2}grams "). The success of the alignment is related to the similarity between chromatograms. The 'inner-product' correlation was used successfully for this purpose. The average inner-product correlation of the five "new" chroma^{2}grams with the 24 "old" chroma^{2}grams in the dataset was 0.06 prior to the transformation. After transformation of the data, this inner-product correlation had been increased to 0.74.
The effect of the transformation was evaluated by PCA (on the linear, modulated signal) and by Parafac (on the demodulated matrix). Although the score plots obtained by the two techniques showed much resemblance, the percentage of variance captured in the first PC (from PCA) or factor (from Parafac) was 48.0% and 32.0%, respectively, before transformation and 80.3% and 56.8%, respectively, after transformation.
The reported approach left the quantitative chromatographic information (peak areas) essentially unchanged, which is a very important requirement.

The same approach was applied to four total-ion-current chroma^{2}grams recorded by a GC×GC-TOF-MS at different flow rates, the inner-product correlation was found to increase from 0.11 to 0.77 upon transformation of the data. The first principal component from PCA captured 58% of the variance in the original data, whereas 93% was captured after transformation of the chromatograms.

The Parafac2 method proved capable of modeling the unaligned GC×GC data and the results were very similar to those obtained when the conventional Parafac method was applied to the aligned data. However, Parafac2 did require a substantial computational effort. Yet, since it eliminates the necessity for an alignment step, Parafac2 may be a serious option for the multivariate analysis of comprehensive chroma^{2}grams. The percentage variance captured in a two-factor model does not significantly differ before and after transformation (69.3% and 68.2%, respectively), demonstrating that alignment is not needed in conjunction with the Parafac2 method.

Unfortunately, the direct comparison of the factor scores between Parafac and Parafac2 is not possible, because the two methods require different input data. The image-processing tools used in this study are limited to components that appear in the chroma^{2}grams and data matrix at their "real" second dimension retention times. Component peaks that arise from a following or preceding modulation exhibit a smaller or larger time shift. The only way to establish the appropriate shifts for all components is to calculate 'real' second dimension retention times ("dewrapping"). This is not just a drawback for the proposed method, but for any method that employs the "demodulated" data matrix (chroma^{2}gram).

The present study was conducted on relatively "clean" samples using relatively mild temperature programs. Similar procedures are yet to be applied on very complex samples and when using temperature programs approaching or exceeding the maximum operating temperature of the columns. However, the results obtained so far are highly encouraging and this suggests that the further study and application of image-processing tools for peak-alignment in GC×GC may be very worthwhile.

Acknowledgements

The authors would like to thank Maud Koek (TNO Nutrition and Food Research) for the GC×GC-TOF-MS data.

Chapter 6

Classification of crude oils with GC× GC and multivariate techniques*

Comprehensive two-dimensional gas chromatography (GC×GC) has proven to be an extremely powerful separation technique for the analysis of complex volatile mixtures. This separation power can be used to discriminate between highly similar samples. In this Chapter we will describe the use of GC×GC for the classification of crude oils from different reservoirs within one oil field. These highly complex chromatograms contain about 6000 individual, quantified components. Differences between reservoirs only manifest themselves by small differences in the levels in most of the 6000 components. For this reason, multivariate analysis (MVA) techniques are required for finding chemical profiles describing the differences between the reservoirs. Unfortunately, such methods cannot discern between 'informative variables', *i.e.* peaks describing differences between samples, and non-informative variables', *i.e.* peaks not describing relevant differences. For this reason, variable-selection techniques are required. A selection based on information between duplicate measurements was used. With this information, 292 peaks were used for building

*Submitted for publications as: *Classification of highly similar crude oils using data sets from comprehensive two-dimensional gas chromatography (GC×GC) and multivariate techniques*, V.G. van Mispelaar, A.K. Smilde, O.E. de Noord, J. Blomberg and P.J. Schoenmakers, Journal of Chromatography A

a discrimination model. Validation was performed using the ratio of the sum of distances between groups and the sum of distances within groups. This step resulted in the detection of an outlier, which could be traced to a production problem, which could be explained retrospectively.

6.1 Introduction

Chemists working in (gas) chromatography are continuously faced with improved instrumentation and techniques. Developments in injection techniques facilitate the injection of large volumes and 'dirty' samples, while selective detection allows detection of components at low levels. Moreover, developments in electronics, such as flow control, strongly improve the repeatability and reproducibility of the technique. All these developments have resulted in a dramatically enhanced robustness of (gas) chromatographic methods. They also create the possibility to analyze large numbers of samples in a more-or-less automated way, facilitating other types of applications, such as high-throughput analysis and metabolism studies (metabolomics).
Instrumental advances also affect the applicability of comprehensive two-dimensional gas chromatography. With this technique, highly complex volatile mixtures can be analyzed in unsurpassed detail. GC×GC can separate complex samples into thousands of individual components. Most examples in the literature concern a single or a few samples. However, the comparison of a series of GC×GC chromatograms (or chroma^{2}grams) can yield very valuable information as well. Especially for highly similar samples, high-resolution techniques are essential to reveal minute differences.
Large datasets require other processing approaches than conventional chromatograms. If there is no prior information regarding components of interest, the traditional approach of quantifying all components present and comparing them univariately is clearly not an attractive strategy. MVA techniques provide better options for processing such large datasets. Such an approach is already adopted in, for example, the field of metabolomics [74, 120]. By comparing two (or more) groups of subjects (*e.g.* sick *vs.* healthy, treated *vs.* untreated), valuable information on metabolic differences between these groups can be obtained. However, this information can only be attained when the number of objects is sufficiently large to eliminate natural variation between the subjects. This approach is not restricted to systems biology.

It is also is applicable to other highly complex mixtures, such as crude oil. The chemical composition of crude oil is determined by its origin and geochemical history. Both chemical composition and boiling-point range can vary widely between different oil fields. However, different reservoirs within one field have a similar origin and a highly similar geochemical history, which can result in minute differences in chemical composition.

Crude oil can contain hydrocarbons from C_4 up to C_{100} or even higher, and the number of theoretical isomers is stunningly large. Techniques such as GC×GC and GC×GC-TOF-MS are by far not sufficient to reveal the full complexity of this class of mixtures. The number of components that can be separated and identified using these techniques is nonetheless impressive.

From a chemometric point of view, these chromatograms are highly interesting. Each object (or sample) is described by a large number of variables (or peaks). Classification of these objects according to their origin can be achieved using discriminant-analysis (DA) methods. Such methods try to find profiles of variables in the data that differentiate between groups of objects. *A priori* information (which object belongs to which group) is required. In many cases this information seems obvious. For example, patients are healthy or sick. However, this information is not necessarily correct. In the example used, patients may not be diagnosed correctly or they may be suffering from other disorders. Incorporating incorrect information into these so-called "supervised techniques" will clearly lead to erroneous models. On the other hand, exploratory techniques, such as principal component analysis, are not beneficial if the data contains a high number of non-informative variables. Therefore, a combination of supervised techniques, for the discovery of discriminating variables, and unsupervised techniques, for finding natural clusters in data, is potentially very strong.

In this study we will apply GC×GC to a set of crude-oil samples from three reservoirs within an oil field. Since no prior knowledge was available on the chemical components that would discriminate between the three fields, as many components as possible needed to be separated and quantified. Multivariate-analysis techniques facilitated the recovery of discriminating components or component profiles.

6.2 Theory

6.2.1 GC×GC

One of the greatest and most significant advances for the characterization of complex mixtures of volatile compounds has been the advent of comprehensive two-dimensional gas chromatography (GC×GC). This technique was pioneered and advocated by John Phillips [1–3]. Two different GC columns are used in GC×GC. The fist-dimension column is (usually) a conventional capillary GC column, with a typical internal diameter of 250 μm. Most commonly, this column contains a non-polar stationary phase, so that it separates components largely based on their vapour pressures (boiling points). The second-dimension column is considerably smaller (smaller diameter, shorter length) than the first-dimension column, so that separations in the second dimension are essentially much faster. The stationary phase is selected such that this column separates on properties other than volatility, such as molecular shape or polarity. The two columns are coupled using a so-called modulator. This device facilitates the continuous accumulation, refocusing and injection of small portions of the first-column effluent into the second-dimension column. With each modulation, a new second-dimension chromatogram is started. The detector, which is positioned at the end of the second-dimension column, records these fast chromatograms. At the end of a chromatographic run, the chromatogram contains many of these fast separations in series. After 'demodulation' [5], a chroma^{2}gram is obtained, which is usually represented by a colour or contour plot. In many applications, GC×GC has proven to be an excellent technique for the separation of very complex samples, such as petrochemical products [64, 77, 107] essential oils [60, 108], fatty acids [68, 69], doping control [61], flavour analysis [62], residue analysis [63], and cigarette smoke [114].

6.2.2 Data analysis

Many analytical techniques exist that can generate large datasets. The human mind is only capable of interpreting data in three-dimensions. Visualization of higher-dimensionality data requires reduction techniques. Fortunately, MVA offers various approaches to reduce the data dimensionality. Classification and clustering problems can be solved using two types of

techniques. *Exploratory methods* extract (natural) patterns from the data. *Supervised classification techniques* use prior information (which objects belong to which groups) to find differences (or similarities) between groups of samples.

Exploratory methods

PCA

The most commonly encountered exploratory method is principal-component analysis (PCA). In PCA, the original variables are replaced by a (strongly) reduced number of uncorrelated (orthogonal) variables, called the principal components. Mathematically:

$$\boldsymbol{X} = T \times P^T + \boldsymbol{E} \tag{6.1}$$

Where:

$\boldsymbol{X}$	Original dataset containing n (samples)$\times$ p (variables)
T	scores of n (samples) $\times$ F (principal components)
P^T	transposed loadings containing F (principal components) $\times$ p (variables)
$\boldsymbol{E}$	Residuals, *i.e.* variation not explained by the model

The principal components are constructed in such a way, that the first principal component (PC1) represents the main source of variation in the original dataset. The second PC is orthogonal to the first and represents the maximum variance not explained in PC1. Each PC is a linear combination of the original variables. The position of each variable is expressed in the principal-component loadings. The number of PC's gives an indication of the model complexity. If the data are highly correlated, a few PC's will be sufficient to reproduce the original data. A way of presenting the results obtained using this technique is a score plot. Related objects (belonging to the same groups) have similar scores and will consequently tend to cluster.

Projection Pursuit

Another unsupervised projection technique is Projection Pursuit (PP) [121]. Unlike PCA, the main objective of which is to explain variance in the data, PP searches interesting low-dimensional linear projections in the data. This is achieved by optimizing the projection index, which can be regarded as

an objective function. In literature, several projection indices have been described [121].

Supervised techniques

Discriminant-analysis (DA) [122,123,123] methods can be applied if attention is focused on differences between known groups of samples. The technique is based on the assumption that samples from the same group are more similar than samples belonging to different groups. The goal of DA is to find and identify structures in the original data, which show large differences between the group means. This process requires *a priori* knowledge on which samples belong to the same group. Discriminant analysis has been used for a wide variety of problems in analytical chemistry. For example, the differentiation of coffee [124, 125], wine [31, 126, 127], and many other types of samples have been described.

Many discriminant methods have been described in the literature. Both Fischer's linear discriminant analysis FLDA and quadratic discriminant analysis QDA [128] can be used in cases where the number of objects (greatly) exceeds the number of variables. In situations where the number of variables exceeds the number of objects, PCA and partial least squares (PLS) are used to reduce the dimensionality of the data. The principal components or latent variables are then subjected to linear discriminant analysis. These techniques are described in the literature as partial-least-squares discrimination analysis (PLSDA) [129]] and principal-component discriminant analysis (PCDA) [53]. They have been used successfully in various types of applications. Regularized discriminant analysis RDA [130] has been proposed for datasets where the number of variables only slightly exceeds the number of objects.

PCDA

Discriminant analysis (DA) of data containing more variables than objects can be preceded by principal-component analysis (PCA) to reduce the number of variables. The projections (scores) of the samples on the principal components (PC's) are used as a starting point for FLDA. Graphical representation of both the objects (in a score plot) and the discriminant loadings provides valuable information on relations between objects and on important variables in the dataset.

Validation

There are several ways to validate a (discrimination) model. In cross validation one or several objects are excluded, a model is created using the remaining objects, and the group membership of the excluded sample is predicted. A well-balanced model results in a minimal number of false assignments. Another method to assess the validity of a model is by permutation. In this process, the effects of the random assignment of objects to groups are examined.

Figure 6.1 gives a graphical representation of a hypothetical dataset.

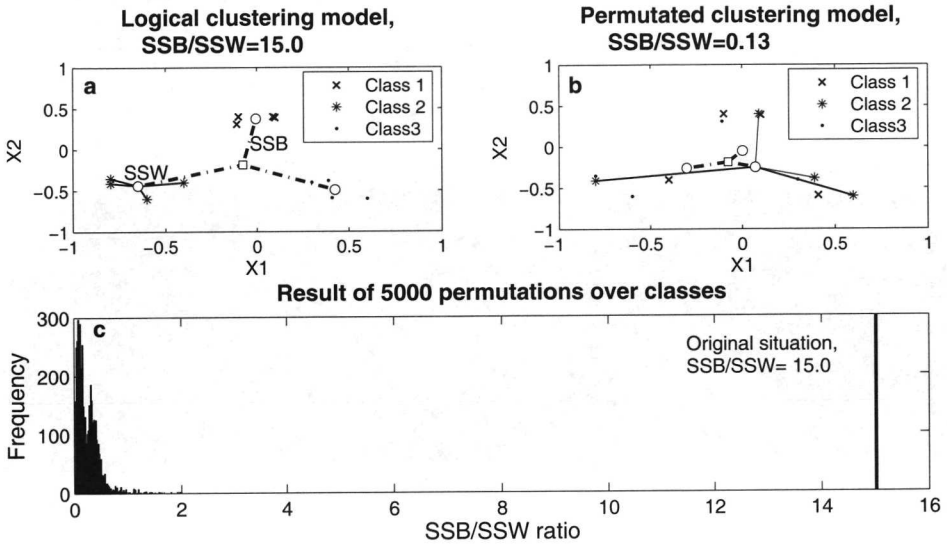


Figure 6.1: Explanation of sum-of-squares within and sum-of-squares between groups.

In this Figure, twelve objects are described by two variables, X1 and X2, located in three groups. Suitable classification of these objects would lead to three dense populations, whereas the distances between the populations should be large. The 'within-group distance' gives a measure for the density of the clusters and can be obtained by calculating the distances for each object to the center of its group. The 'between-group distance' can be used as a measure for the separation of the three clusters and is obtained

by calculating the distance between the group centers. The ratio of the 'sum of distances between groups' and the 'sum of distances within' groups should be maximal for proper clustering. Since the 'sum-of-squares' is used in the calculation, we will refer to SSB and SSW for sum-of-squares distances between and within the groups, respectively. The initial situation in Figure 6.1a results in an SSB/SSW ratio of 15.0. In a permutation process objects are assigned randomly to one of the three groups. The result of a first (random) permutation is shown in Figure 6.1*b*. The sum of distances within the clusters increases significantly, while the sum of distances between the groups changes only slightly. This has a dramatic effect on the ratio SSB/SSW (0.5). Repeating this permutation process many times results in equally many SSB/SSW ratios. A histogram of all these results is presented in Figure 6.1c. Most of the random permutations result in SSB/SSW values between 0 and 1. The original situation, with a SSB/SSW ratio of 15.0 is clearly the best classification of the data.
The above calculations can be described mathematically for the between-group distance [43]:

$$SSB = \sum_{i=1}^{g} m_i \times (\bar{x}_i - \bar{x})^2 \tag{6.2}$$

For the within-group distance:

$$SSW = \sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i) \tag{6.3}$$

Where

g	Number of groups
m	Number of objects for group i
x_{ij}	Object i of group j
$\bar{x}_i$	Mean of group i
$\bar{x}$	Overall mean of $\bar{x}_i$

6.3 Experimental

Instrumentation The samples were analyzed using an Agilent 6890 GC (Wilmington, DE, USA), equipped with a CTC CombiPAL autosampling

unit (CTC Analytics, Zwingen, Switzerland), and a CIS-4 Programmed-Temperature-Vaporization (PTV) injector (Gerstel, Mulheim an der Rühr, Germany). This system was retrofitted with a Zoex KT2003 thermal modulator and equipped with a second dimension-column oven (Zoex, Lincoln, NE, USA), enabling independent second-dimension column heating.
The column-set consisted of a 10 m length × 250 μm internal diameter × 0.25 μm film thickness DB-1 column (J&W Scientific, Folsom, CA, USA) in the first dimension and a 2 m length 0.1 mm internal diameter × 0.1 μm film thickness BPX50 column (SGE, Ringwood, Australia) in the second dimension. The modulation was performed in a 1.6 m × 0.1 mm diphenyltetramethyl-disilazane (DPTMDS) deactivated fused silica capillary (BGB Analytik, Anwil, Switzerland). A fused-silica capillary of the same material with a length of approximately 50 cm was used to connect the second-dimension column to the flame-ionization detector (FID). Columns were coupled with custom-made press-fits (Techrom, Purmerend, The Netherlands). The carrier gas was helium at a constant head pressure of 250 kPa, resulting in a column flow of approximately 1 mL/min at 40°C.
The temperature for the first-dimension column oven was programmed from 40°C (5 minutes isothermal) at a rate of 2°C/min to 300°C (20 minutes isothermal), followed by a negative ramp of 13°C/min to 40°C (10 minutes isothermal). Both the hot pulse of the release jet and the secondary oven were operated at an offset of 50°C above the temperature of the primary oven.
The PTV injector was programmed from 40°C to 250°C (5 minutes isothermal) with a ramp of 12°C/s. These conditions resulted in a selective discrimination from C_{30} hydrocarbon upwards. The modulation time was 7.5 s and the hot-pulse duration was 300 ms. Liquid-nitrogen-cooled nitrogen gas was used as modulating agent at a flow of 17 L/min. A Zoex auto-fill unit was used to enable continuous operation.

6.3.1 Instrument control and data processing

Instrument control and data acquisition were achieved with EZ-Chrom Elite (v2.61, SSI, Willemstad, the Netherlands). Data were collected at 100 Hz to obtain a sufficiently large number of datapoints across a peak. Chromatograms were exported to the Common Data Format (CDF). Data han-

dling was performed in MATLAB R14 (The Mathworks, Natick, MA, USA) running on a Compaq EVO W6000 computer equipped with 1 Gb of RAM. Data-handling routines were developed in-house. In addition, the NetCDF toolbox [104] was used.

6.3.2 Samples

A set of 14 different oil samples, originating from one oil field, was selected by Shell International Exploration and Production (SIEP, Rijswijk, The Netherlands). The samples were divided in the three subclasses *A*, *B* and *C*, referring to the reservoirs within the original oil field. The samples were diluted ten-fold in cyclohexane (*p.a.* quality, Merck, Darmstadt, Germany) containing 0.1% (*w*/*w*) 1,2-dichlorobenzene (*p.a.* quality, Merck) as an internal standard.
All samples were analyzed in duplicate. One sample was analyzed in fivefold. In the sequence two blanks and an alkane mixture containing C_5-C_{42} hydrocarbons in CS_2 were included.

6.4 Results and discussion

Samples were analyzed in one sequence in order to reduce retention variations. The duration of the negative ramp in the oven program was controlled to obtain a highly repeatable temperature program, thereby reducing retention-time shifts. The alkane mixture was used to "spline" the data (described in Section 4.2.3, on page 61 of this thesis). In this process *n*-alkanes were shifted so as to obtain constant second-dimension (relative) retention times. The peak positions for a homologous series of *n*-alkanes were used to create a piecewise-linear shift function. This function was subsequently applied to all samples in the sequence.
Since the data were acquired directly from the FID, a series of second-dimension chromatograms was registered. Integration of this signal therefore resulted in integrated second-dimension peaks. However, the modulation process typically resulted in three or four modulations across a first-dimension peak. The total peak area of a certain component was the sum of areas in the different modulations. The automated integration in the first dimension direction was performed with an algorithm developed in-house.

In this step the number of peak positions found was typically reduced by a factor of 2.5. Figure 6.2 shows the chroma^{2}gram of a typical crude-oil sample.

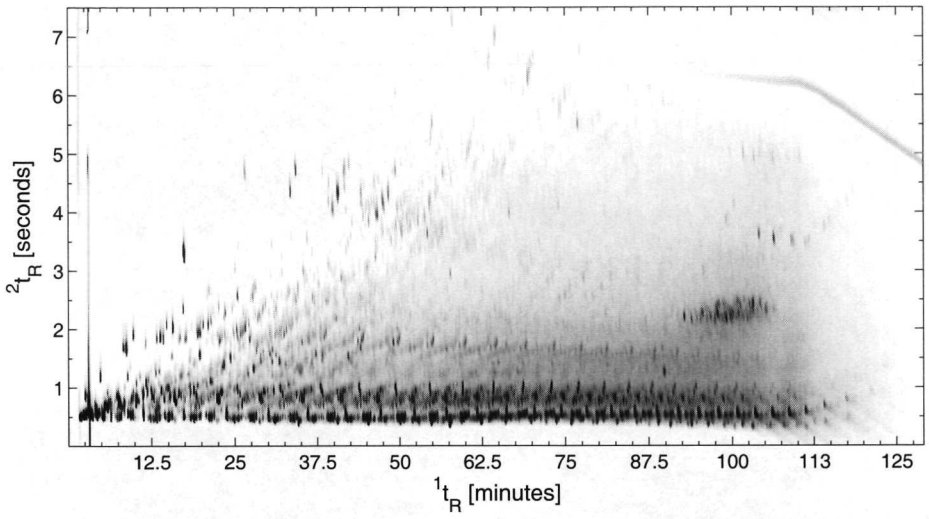


Figure 6.2: Typical crude oil chroma^{2}gram.

6.4.1 Pre-processing

Typical crudes can contain about 6000 individual integrated peaks. This number includes peaks eluting in the isothermal region of the chromatogram and components that are not interesting for quantitative analysis due to the selective discrimination of the PTV. Components eluting from a first-dimension retention time of 106 minutes upwards were not quantitatively transferred from the injector to the column and therefore eliminated.

Alignment

Unfortunately, chromatographic techniques suffer from retention-time shifts. This results in inconsistent retention time within a series of samples. Since MVA techniques are unable to deal with shifting peaks, alignment is a required pre-processing step. An alignment routine developed in-house was

used to eliminate small variations in peak-apex locations. The maximum allowed shift was one data point in the first dimension (= 7.5 seconds) and 15 points in the second dimension (=0.15 seconds). Figure 6.3 shows the position of the aligned peaks.

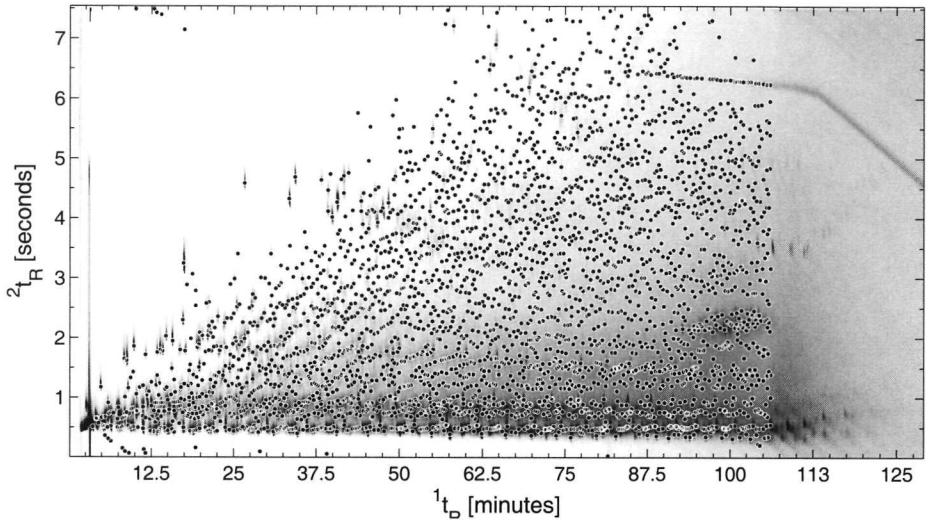


Figure 6.3: Location of peaks after alignment.

After cut-off and alignment, a selection of 3904 peaks remained. The resulting dataset contained 30 × 3904 (objects or samples × variables or peaks). Such well-described data should (at least theoretically) be very suitable for multivariate-analysis techniques. However, the PCDA result after 'mean-centering' was disappointing (Figure 6.4).

A good classification model should form dense, separated clusters. In our initial situation PCDA resulted in overlapping clusters, indicating no separation between groups. The SSB/SSW plot (Figure 6.5) also turned out to be highly unsatisfactory. The proposed classification turned out to be no better than a random classification. This observation can partly be explained by 'over-fitting'. Since each object is described by 3904 variables, the number of objects should be much larger than 30 to obtain proper classification results. This seems trivial, but it is a very important problem within the MVA field. Many analytical techniques are able to provide highly detailed information, resulting in large sets of data.

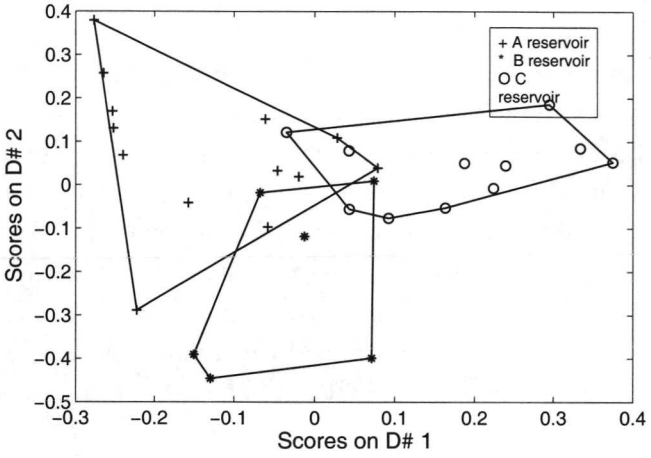


Figure 6.4: PCDA of 3904 aligned peaks in 30 objects.

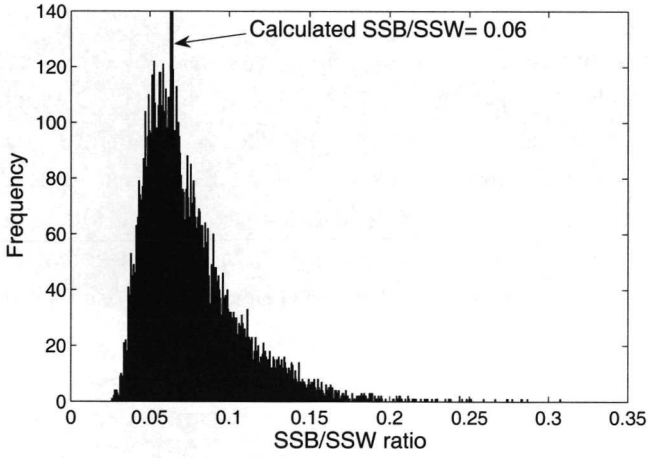


Figure 6.5: SSB/SSW distribution of 1000 random permutations.

The number of available samples usually does not increase accordingly. Even dimension reduction using PCA was insufficient to abstract sufficient relevant information, despite the captured variance in 10 PC's being 83%. The second problem is the presence of non-informative data, which is irrelevant for the differentiation between reservoirs. Many peaks are indeed

of little or no relevance. They contribute hardly or not to the desired discrimination model. The other source of irrelevant data is the integration process. Integration of highly complex chromatographic signals inevitably results in errors. Baseline-separated peaks can be quantified very accurately; convoluted peaks are much-more difficult to integrate. In the case of crude oil, certain regions of the chromatogram do not contain any baseline, due to the continuous elution of components. Quantification of such a signal obviously does not yield relevant data, since the integration errors obscure relevant information.
However, these irrelevant data are included when building the discrimination model and performing the SSB/SSW calculation. Distinction between informative and non-informative peaks can be achieved by variable selection.

6.4.2 Multivariate analysis

Variable selection

Variable selection is commonly performed to (strongly) reduce the number of variables in a dataset. However, many variable-selection strategies can be considered to be supervised, *i.e.* variables are identified which support a certain group structure. Such supervised selection routines aim at finding components supporting the proposed classification structure. However, a different classification structure will also lead to a selection of components. These routines are therefore solely dependent on the classification structure. Even in a dataset containing random numbers, supervised classification is capable of selecting a number of (random) variables that would support the classification. Unsupervised variable selection seems, therefore, to be a more appropriate choice. A suitable criterion can be established by using the information gained from duplicate measurements. Well-separated peaks in duplicate measurements show small relative standard deviations (r.s.d.). In our situation, there were 14 samples and thus as many r.s.d. values for each of the 3904 peaks. A selected peak should have a small r.s.d. value for each of the 14 samples. The average r.s.d. value for any component over all the 14 samples should be as low as possible. Also, the standard deviation between the 14 r.s.d. values should be minimized, excluding variables for which one of the samples has a large r.s.d., while other objects have a small r.s.d. value. By restricting the average r.s.d. between duplicate

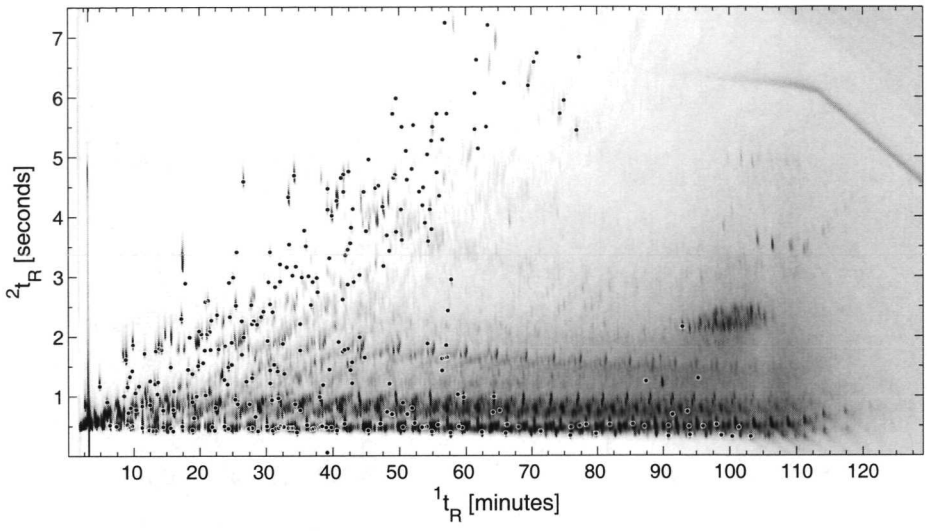


Figure 6.6: Peaks selected on r.s.d between duplicates.

measurements (for each of the 3904 peaks) to 10%, 292 variables were selected. Figure 6.6 shows the position of the selected peaks. Subsequent PCDA revealed clustering according to the reservoir origin. Inspection of the DA loadings did not reveal any specific 'biomarker components' that could be used to discriminate between reservoirs. Differences between the three reservoirs were the result of many small differences between the 292 selected peaks.

Manual selection

Samples from the different reservoirs could not be discriminated based on one or a few components (so-called biomarkers). Rather, the differences in all of the included peaks had to be considered. Therefore, verification of the groups had to be performed using unsupervised MVA techniques applied to a small subset of the data. To this end, a selection of 65 peaks was manually extracted from the chromatograms. The selection consisted of baseline-separated components. Figure 6.7 shows the peak positions.

The resulting dataset was significantly better defined, having the dimension of 30×65 (samples $\times$ peaks). Before data-analysis, mean centering was applied as a pre-processing technique. Figure 6.8 shows the result of

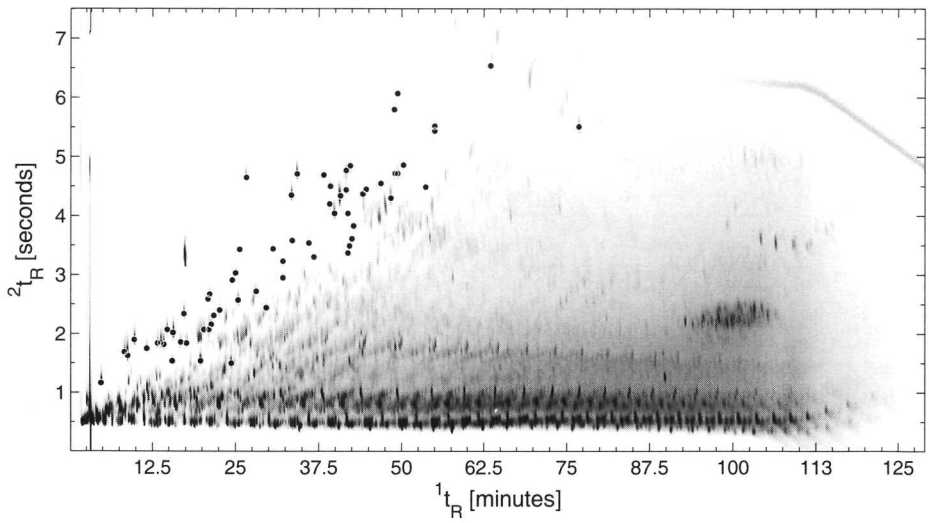


Figure 6.7: Position of 65 manual selected peaks.

principal-component analysis. With only two PC's, 96.9% of the variance was captured. Samples in all groups (A, B and C) formed dense clusters, implying a high similarity between the members of each group. However,

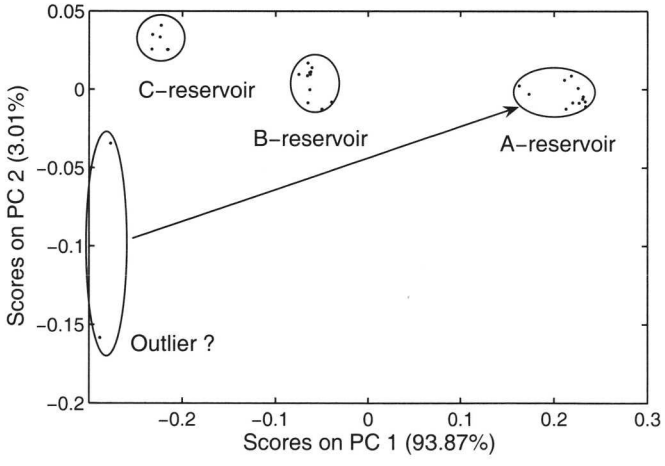


Figure 6.8: PCA after mean centering of 65 manually selected peaks.

both duplicates of sample 4S94A seem to be very different from the other *A*-group members. These samples must therefore be considered as outliers. Based on these results, the two samples are likely to belong to a different group (*e.g.* originate from a different reservoir). Calculation of SSB/SSW

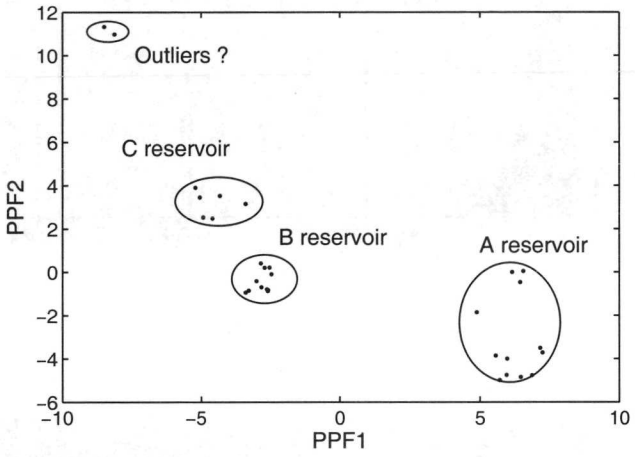


Figure 6.9: Projection pursuit after mean-centering.

values can numerically support the outlier hypothesis. The distance between the two duplicates of 4S94A can be explained by the small percentage of variation in PC2. Small differences between the samples are blown out of proportion.Projection pursuit yielded a somewhat improved clustering results, as shown in Figure 6.9.The observation that two samples are not classified correctly obviously has severe implications for discriminant analysis. Grouping or clustering of samples of incorrect origin evidently results in the calculation of incorrect DA-loadings and scores. There are a number of possible solutions for this problem. The first is to simply remove the two samples from the dataset before the PCDA step. Trying to fit the samples into one of the two other groups may also be a possible solution. The third option is to define a new group in which the two duplicates of the samples are included. These three options were all investigated. Results can be found in Figures 6.10 and 6.11. The SSB/SSW ratios were dramatically improved, while the discriminant analysis resulted in much denser clusters. Based on these results, sample 4S94A is best described as a new group, since this hypothesis results in the best (highest) SSB/SSW ratio. The

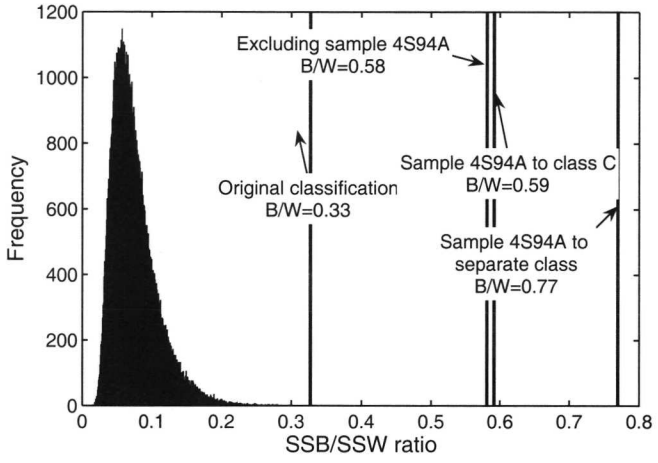


Figure 6.10: SSB/SSW results of 1000 random permutations.

suppliers of the samples gave the ultimate proof for the hypothesis that sample 4S94A was different from the other A-group members. This specific sample was taken during a pipeline leakage. Instead of producing oil from

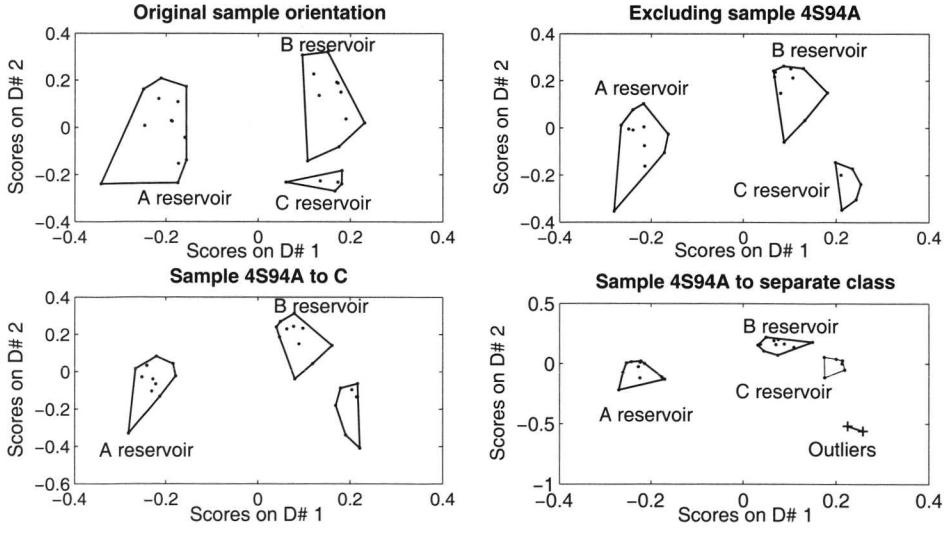


Figure 6.11: PCDA results of different scenarios.

the A-reservoir, a mixture of A and C was produced. The third option in Figures 6.10 and 6.11. describes this situation best. However, this conclusion is not in line with the PCA results. Since PCA scores are a linear combination, mixtures of the groups should fall in between the pure groups. This may be explained by the small number of components (65 out of 6000) taken into consideration. These peaks might not be representative for the entire sample.

6.5 Conclusions

Comprehensive two-dimensional gas chromatography proved to be highly suitable for measuring small differences between complex samples. This has already been demonstrated in various applications in the literature. However, improved modulation techniques (*e.g.* cryogenic modulation) lead to a drastically improved stability of retention times. This facilitates the comparison of large series of samples and the use of sophisticated (multivariate) data-analysis methods.

The measured samples, consisting of crude oils from three reservoirs within one oil field, were highly similar. The necessary pre-processing techniques, such as integration and alignment, resulted in 3904 peaks found in all 30 samples. However, using discriminant analysis on this dataset, we were unable to classify the samples according to their origin. Variable selection turned out to be essential to eliminate the problem of over-determination of the data matrix. Selection of variables based on the average relative standard deviation between duplicate measurements, with an upper limit of 10%, resulted in a reduction to 292 variables (peaks). When these data were subjected to PCDA, separate clusters describing the different reservoirs were obtained.

Verification of the groups with PCA and PP resulted in the discovery of an outlier. Feeding this information into PCDA did improve the results dramatically.

The validation with the ratio of SSB/SSW (sum-of-squares-between-groups to sum-of-squares-within-groups) proved unambiguously that the proposed classification was superior to the original classification. This supported the hypothesis that the initial classification structure contained an incorrect entry. The outlier in the dataset could be explained retrospectively from

production problems.

Acknowledgements
The authors would like to acknowledge Florian Wülfert for many stimulating discussions on classification techniques.

Summary

In this thesis the combination of multivariate analysis (MVA) or chemometrics and two-dimensional comprehensive gas chromatography (GC×GC) is described. The sheer complexity of the two-dimensional chromatograms (chroma^{2}grams) is the most important incentive for pursuing this combination. The title of the thesis is 'Chromametrics', which is a contraction of chromatography and chemometrics.

In **Chapter 1** the combination of chemometrics and chromatography is put in perspective.

In **Chapter 2**, a classification scheme for applications of (gas) chromatography is presented. Almost all chromatographic applications can be classified into a small number of well-defined categories.
The first identified category is 'Target-compound analyses', in which concentrations of a limited number of pre-identified components is desired.
The second category is identified as 'Group-type analyses'. These applications focus on the quantification of groups of components. Often these groups have structural properties in common, such as aromatic rings, double bonds, *etc.*.
The last identified group of applications is referred to as 'Fingerprinting'. In such applications, correlations between analytical composition data (the chroma^{2}gram) and product properties are established. Classifying (almost) applications of chromatography into the three aforementioned groups aids the developers on the forefront of technology to judge the merits of new developments. Practical users can use this scheme to decide on the applicability of new developments for their specific application. The applicability of this scheme is demonstrated on GC×GC as such and on its

combination with MVA. **Chapters 3** to **6** deal with examples of each of the proposed generic application types.

In **Chapter 3** the use of so-called multiway methods, such as parallel-factor analysis (Parafac) and multilinear partial-least-squares (NPLS) is described for the quantification of a limited number of components in complex samples. In this example, synthetic perfume mixtures are used to demonstrate the strengths of the *Chromametrics* approach. Compared to integration, which is considered to be a benchmark technique, multiway methods perform only slightly worse with respect to accuracy and precision. With respect to speed, multiway methods are far superior to integration.

Chapter 4 describes chemometric tools for group-type analysis by GC×GC. In this application, the total levels of component groups with identical structural characteristics. Generic data-handling steps for obtaining a good chroma^{2}gram are described. In addition, tools for enhancing the visualization (*e.g.* baseline correction and splining) are described. A possible group-wise integration strategy is described. Branched alcohols are used to illustrate this approach. A route towards the elimination of retention-time shifts in this type of applications is also described.

In **Chapter 5** an alignment technique for two-dimensional separation techniques is demonstrated. This approach uses a second-order polynomial transformation profile to align the entire chroma^{2}gram. Such comprehensive alignment techniques are essential when highly detailed chroma^{2}grams are combined with chemometric techniques in fingerprinting applications. The applicability of the alignment technique is demonstrated for GC×GC with flame-ionization detection (FID) as well as for GC×GC coupled with time-of-flight mass spectrometry. Multivariate techniques (principal-component analysis (PCA) and Parafac) are used to assess the quality of the alignment. Alternatively, Parafac2 is used as a way to deal with retention-time shifts within the multivariate method without the need for prior alignment.

In **Chapter 6**, the use of GC×GC data for the classification of crude oils is described. Variable-selection techniques were required to discriminate between 'informative' and 'non-informative' data. Both supervised and

unsupervised classification techniques were used to detect an outlier in the samples. The samples were successfully classified according to their reservoir origin.

Samenvatting

In dit proefschrift wordt de combinatie van multivariate analyse (MVA) technieken, ofwel chemometrie, en twee-dimensionale scheidingsmethoden ("comprehensive" twee-dimensionale gas chromatografie).
De belangrijkste reden om dit te doen is de geweldige complexiteit van de data. De titel 'Chromametrics' is een samentrekking van *chromatography* en *chemometrics*.

Hoofstuk 1 geeft een kort historisch perspectief van de combinatie van chemometrie en chromatografie.

Hoofdstuk 2 beschrijft een classificatie-schema voor (gas)chromatografische scheidingsmethoden. Met dit schema kunnen vrijwel alle gas-chromatografische toepassingen worden ingedeeld in slechts drie generieke applikatie typen. De drie gedefinieerde applikaties zijn de 'Doelkomponenten analyze', 'Groep-type analyse' en 'Fingerprinting'. Het eerste type richt zich op het achterhalen van concentraties van een beperkt aantal verbingen in een complexe matrix. Het tweede type richt zich op het achterhalen van concentraties van groepen van verbindingen. In het algemeen hebben de componenten binnen een groep een overeenkomst in chemische structuur, of gedragen ze zich identiek bij chemische omzettingen. De laatste klasse van toepassingen, het 'Fingerprinten', richt zich op het leggen van correlaties tussen het twee-dimensionale chromatogram en de eigenschappen van een bepaald produkt. Op deze manier kunnen verbingen die zorgen voor het slecht presteren van een product worden achterhaald.
Het classificatie-schema moet zowel de ontwikkelaars van nieuwe toepassingen om de voordelen van hun verbeteringen voor de praktische gebruikers van nieuwe toepassingen. Aan de andere kant kunnen de

gebruikers van deze toepassingen eerder inzien wat de voordelen van een nieuwe toepassing voor hun inhoud. De bruikbaarheid van het schema wordt gedemonstreerd op GC×GC als een op zichzelf staande techniek, en in combinatie met MVA.
De **Hoofdstukken** 3 tot 6 gaan over de beschreven generieke toepassingen.

Hoofdstuk 3 beschrijft het gebruik van zogenaamde meerweg-technieken, zoals Parafac en NPLS, voor het berekenen van concentraties van een beperkt aantal verbindingen in een complex monster. In dit voorbeeld zijn synthetische parfums gebruikt om de voordelen van deze benadering te illustreren. In vergelijking met integratiemethoden, die in dit geval beschouwd worden als referentie, presteren meerweg-technieken slechts marginaal slechter op nauwkeurigheid en precisie. Als gekeken wordt naar de snelheid, dan zijn meerwegmethoden verreweg superieur aan de traditionele methoden.

Hoofstuk 4 beschrijft de chemometrische methoden voor de analyze van z.g. groep-typen. In deze toepassing is men geinteresseerd in het achterhalen van de concentraties van stoffen met eenzelfde karakteristiek (structuur, gedrag bij bepaalde omtzettingsomstandigheden *etc.*). Een aantal algemene stappen om een chroma^{2}gram te verkrijgen worden beschreven. Daarnaast worden een aantal technieken berschreven die de visualizatie ten goede komen (het corrigeren voor de basislijn en 'splining' van de datamatrix). Ook wordt een mogelijke aanpak voor het berekenen van de concentraties van groepen van verbindingen beschreven. De voorbeelden worden geillustreerd aan de hand van monsters waarin vertakte alcoholen zitten. Een mogelijke aanpak van de verschuivingen in retentietijden wordt ook beschreven.

Hoofdstuk 5 beschrijft de toepassing van een techniek om verschillen in retentietijden in twee dimensies op te lossen. Deze toepast maakt gebruik van een tweede-orde vergelijking om een verschuivingsprofiel op te stellen. Op deze manier kan het gehele chroma^{2}gram worden aangepast voor verschillen in retentietijden. Dit is een vereiste omdat vrijwel alle chemometrische technieken niet met deze verschillen kunnen omgaan. Zeker voor het gebruik van GC×GC voor de 'Fingerprint' toepassing is dit

noodzakelijk.
De techniek is toegepast op GC×GC met een vlamionisatie detector (FID) en met een vlucttijd massaspectrometer (TOF-MS).
Multivariate technieken zoals PCA en Parafac zijn gebruikt om de effectiviteit van deze aanpak te illustreren. Daarnaast is Parafac2 gebruikt als MVA techniek die de verschuivingen in retentietijden vanuit het algoritme aanpakt.

Hoofdstuk 6 beschrijft het gebruik van GC×GC data voor het classificeren van ruwe olie. Selectie technieken om onderscheid te maken tussen relevante en irrelevante data zijn ook toegepast. Zowel de 'gestuurde' als 'ongestuurde' classificatie methoden zijn gebruikt om een verkeerd geclassificeerd monster te achterhalen. Uiteindelijk zijn de monsters geclassificeerd naar hun oorspronkelijke reservoir.

Dankwoord

Een proefschrift schrijf je niet alleen. Van alle betrokken personen wil ik met name de volgende mensen nadrukkelijk bedanken. In de eerste plaats ben ik ontzettend veel dank verschuldigd aan mijn promotor, Peter Schoenmakers. Jij hebt mij ruim vier jaar geleden op het idee gebracht om te gaan promoveren. Mede dankzij jouw altijd scherpe geest is dit proefschrift geworden wat het nu is. Jouw deur stond altijd open en jij had eigenlijk altijd wel tijd voor mij. Ik moet wel toegeven dat werkbesprekingen om acht uur 's morgens zonder koffie geen eenvoudige opgave zijn...
Ik heb veel geleerd van mijn co-promotor, Age Smilde. Het speelse gemak waarmee jij de moeilijkste formules leest en jouw systematische aanpak waarmee jij problemen aanpakt is verbazingwekkend.

Veel ideeën in dit proefschrift zijn gegenereerd in de *Scientific Committee*. De ideen van respectievelijk Peter, Age, Albert, Hans-Gerd, Onno, Jan en Jens, hebben voor een groot deel aan de basis gestaan van de artikelen in dit proefschrift.

Uiteindelijk is dit project mogelijk gemaakt door de bedrijven die dit project samen met TNO zijn gestart. Ik ben hiervoor Arie Meruma (Shell), Hans-Gerd Janssen (Unilever), Robert Jonker (Albemarle) en Nigel Wilson (ICI) zeer erkentelijk. Ze hebben de mogelijkheid geschapen om dit proefschrift te creëren.

Ik ben TNO zeer erkentelijk voor het bieden van de mogelijkheden en de ruimte de laatste vier jaar. Hoewel ik maar een klein gedeelte van die tijd bij TNO ben geweest, heb ik me er zeker thuis gevoeld. Ik heb met veel plezier met Albert mogen samenwerken. Met name jouw onverwoestbare

optimisme ("uitstekend idee, goed werk!") is legendarisch. Florian, jouw kennis en kunde op het gebied van chemometrie is uitzonderlijk (evenals jouw gevoel voor humor). Robert-Jan, onze vergelijkbare situatie (promoveren en aanstaand vaderschap) heeft tot de nodige gespreksstof geleid. Ik zal deze gesprekken zeker missen. Bedankt dat jij mijn paranimf wil zijn. Verder worden Renger, Sabina, Bianca, Jacques, Henk en Elly bedanken voor de gezelligheid en interesse in mijn onderzoek.

Een groot gedeelte van de afgelopen periode heb ik bij Shell doorgebracht. Door gebruik te maken van de apparatuur en faciliteiten heeft dit onderzoek een praktisch stevige basis gekregen. Ik moet hier met name Arie Meruma ("boeiend!") voor bedanken. Verder mag ik de volgende mensen niet vergeten. Ferry, ik heb nog niet eerder een kamergenoot gehad met een slechtere muzieksmaak. Marcel, jouw enthousiasme was aanstekelijk. Verder moet ik Marco, Hassan, Jaap, Piet, Tessa, Henk en Betsie bedanken voor de gezelligheid en interesse. Onno, jouw ideeën over de selectie van variabelen hebben het laatste hoofdstuk een eind op weg geholpen. Jan, de autoritten naar Alkmaar waren misschien niet altijd sneller dan de trein, maar wel een stuk leuker en gezelliger!

Het eerste half jaar heb ik grotendeels doorgebracht op de vakgroep Polymeer-Analyse van de UvA. Ik moet hiervoor (in willekeurige volgorde) Petra, Fiona, Kathalijne, Simona, Maya, Emil, Xulin, Wybren (dat van die stomme Fries neem ik terug...), Remco, Monique, Wim D. Mauro, Rob E.(bedankt, sorry dat het anders liep), Aschwin, Gabriel en iedereen die ik vergeten ben bedanken. Ook na dat eerste jaar hebben jullie ervoor gezorgd dat ik mij er altijd heb thuisgevoeld. Ook de mensen van PAC groep mag ik niet vergeten.

Ik liep al langer met het idee om over te stappen op LaTeX. Het laatste zetje gegeven door *Bill Gates himself*. Zijn fantastische software blijkt uitermate ongeschikt voor het verwerken van complexe documenten. Het was misschien niet de makkelijkste stap, maar het resultaat mag er(hopelijk) zijn. Hans, bedankt voor de hulp hiermee.

Bas, onze vele discussies (varierend van Spaans eten tot de zin van

het bestaan) heb ik altijd erg op prijs gesteld. Dat je mijn paranimf wil zijn, vind ik dan ook erg leuk.

Jan en Nellie, niet in de laatste plaats wil jullie bedanken. Welke keuze ik ook heb gemaakt in mijn leven, ik kon altijd rekenen op jullie onvoorwaardelijke steun.

Els, de afgelopen jaren was ik vaak druk of (geestelijk) afwezig. Zeker naar het einde toe heeft mijn humeur er af en toe flink onder te lijden gehad. Jouw steun en liefde heeft mij erg geholpen. Ik hoop dat wij de komende tijd wat meer tijd voor elkaar hebben. Wouter, jij bent je het zeker niet bewust, maar nog nooit heeft iemand zo'n verpletterende indruk achter gelaten. Met jouw komst is alles een stuk betrekkelijker geworden.

References

[1] J.B. Phillips and Z. Liu. Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface. *Journal of Chromatographic Science*, 29:227–231, 1991.

[2] J.B. Phillips and Z. Liu. Comprehensive multi-dimensional gas chromatography. *Journal of Chromatography A*, 703:327–334, 1995.

[3] J.B. Phillips and J. Beens. Comprehensive two-dimensional gas chromatography: A hyphenated method with strong coupling between the two dimensions. *Journal of Chromatography A*, 856:331–347, 1999.

[4] C.J. Ventrakanami, J. Xu, and J.B. Phillips. Separation orthogonality in temperature-programmed comprehensive two-dimensional gas chromatography. *Analytical Chemistry*, 68:1486–1492, 1996.

[5] J.B. Phillips, D. Luu, Pawliszyn J., and G.C. Carle. Multiplex gas chromatography by thermal modulation of a fused silica capillary column. *Analytical Chemistry*, 57:2779–2787, 1985.

[6] J. Dalluge, J. Beens, and U.A.Th. Brinkman. Comprehensive two-dimensional gas chromatography: A powerful and versatile analytical tool. *Journal of Chromatography A*, 1000:69–108, 2003.

[7] P.J. Marriott, P. Haglund, and R.C.Y. Ong. A review of environmental toxicant analysis by using multidimensional gas chromatography and comprehensive two-dimensional gas chromatography. *Clinica Chimica Acta*, 328:1–19, 2003.

[8] W. Bertsch. Two-dimensional gas chromatography. concepts, instrumentation, and applications. part 1: Fundamentals, conventional two- dimensional gas chromatography, selected applications. *Journal of High Resolution Chromatography*, 22:647–665, 1999.

[9] W. Bertsch. Two-dimensional gas chromatography. concepts, instrumentation, and applications. part 2: Comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 23:167–181, 2000.

[10] J.E. Jackson. *A Users Guide to Principal Components.* John Wiley and Sons, New York, 1991.

[11] H. Martens and T. Naes. *Multivariate Calibration.* John Wiley and Sons, New York, 1996.

[12] J.M. Halket, A. Przyborowska, S.E. Stein, W.G. Mallard, S. Down, and R.A. Chalmers. Deconvolution gas chromatography coupled to mass spectrometry of urinary organic acids: Potential for pattern recognition and automated identification of metabolic disorders. *Rapid Communications in Mass Spectrometry*, 13:279–248, 1999.

[13] C.A. Bruckner, B.J. Prazen, and R.E. Synovec. Comprehensive two-dimensional high-speed gas chromatography with chemometric analysis. *Analytical Chemistry*, 70:2796–2804, 1998.

[14] M. Jalali-Heravi and M. Vosough. Characterization and determination of fatty acids in fish oil using gas chromatography - mass spectrometry coupled with chemometric resolution techniques. *Journal of Chromatography A*, 1024:165–176, 2004.

[15] B.J. Prazen, R.E. Synovec, and B.R. Kowalski. Standardization of second-order chromatographic/spectroscopic data for optimum chemical analysis. *Analytical Chemistry*, 70:218–225, 1998.

[16] C.G. Fraga, B.J. Prazen, and R.E. Synovec. Enhancing the limit of detection for comprehensive two-dimensional gas chromatography using bilinear chemometric analysis. *Journal of High Resolution Chromatography*, 23:215–224, 2000.

[17] J.E. Davis, A. Shepard, N. Stanford, and L.B. Rogers. Principal-component analysis applied to combined gas chromatographic-mass spectrometric data. *Analytical Chemistry*, 46:821–825, 1974.

[18] R.B. Taylor, N.A. Ochekpe, and J. Wangboonskull. Quantitative structure retention relationship studies of some basic antimalarial compounds. *Journal of Liquid Chromatography*, 12:1645–1668, 1989.

[19] R. Kaliszan, K. Osmialowski, B.J. Bassler, and R.A. Hartwick. Mechanism of retention in high-performance liquid chromatography on porous graphitic carbon as revealed by principal component analysis of structural descriptors of solutes. *Journal of Chromatography A*, 499:333–344, 1990.

[20] A. Detroyer, V. Schoonjans, F. Questier, Y. Vander Heyden, A.P. Borosy, Q. Guo, and D.L. Massart. Exploratory chemometric analysis of the classification of pharmaceutical substances based on chromatographic data. *Journal of Chromatography A*, 897:23–36, 2000.

[21] J. Raymer, D. Wiesler, and M. Novotny. Structure-retention studies of model ketones by capillary gas chromatography. *Journal of Chromatography A*, 325:13–32, 1985.

[22] L.A. Currie, J.J. Filliben, and J.R. DeVoe. Statistical and mathematical methods in analytical chemistry. *Analytical Chemistry*, 44:497R–512R, 1972.

[23] S. Wold, E. Johanbsson, E. Jellum, I. Bjornson, and R. Nesbakken. Application of simca multivariate analysis to the classification of gas chromatographic profiles of human brain tissues. *Analytica Chimica Acta*, 133:251–259, 1981.

[24] F.I. Onuska, A. Murdoch, and S. Davies. Application of chemometrics in homolog specific analysis of polychlorinated biphenyls. *Journal of High Resolution Chromatography*, 8:747–754, 1985.

[25] M. Chien. Analysis of complex mixtures by gas chromatography/mass spectrometry using a pattern recognition method. *Analytical Chemistry*, 57:348–352, 1985.

[26] W.J. Dunn, D.L. Stalling, T.R. Schwartz, J.W. Hogan, J.D. Petty, E. Johansson, and S. Wold. Pattern recognition for classification and determination of polychlorinated biphenyls in environmental samples. *Analytical Chemistry*, 56:1308–1313, 1984.

[27] D.S. Lee, B.S. Noh, S.Y. Bae, and K. Kim. Characterization of fatty acids composition in vegetable oils by gas chromatography and chemometrics. *Analytica Chimica Acta*, 358:163–175, 1998.

[28] B. Tan, J.K. Hardy, and R.E. Snavely. Accelerant classification by gas chromatography/mass spectrometry and multivariate pattern recognition. *Analytica Chimica Acta*, 422:37–46, 2000.

[29] B.K. Lavine, A.J. Moores, H.T. Mayfield, and A. Faruque. Fuel spill identification by gas chromatography - genetic algorithms and pattern recognition techniques. *Analytical Letters*, 31(15):2805-2822, 1998.

[30] B.K. Lavine, A.J. Moores, and A. Faruque. Genetic algorithms applied to pattern recognition analysis of high-speed gas chromatograms of aviation turbine fuels using an integrated jet-a/jp-8 databases. *Microchemical Journal*, 61:69–78, 1999.

[31] W.O. Kwan and B.R. Kowalski. Correlation of objective chemical measurements and subjective sensory evaluations. wines of vitis vinifera variety 'pinot noir' from france and the united states. *Analytica Chimica Acta*, 28:356–359, 1980.

[32] A.M. Costa Freitas, C. Parreira, and L. Vilas-Bonas. The use of electronic aroma sensing device to assess coffee differentiation - comparison with spme gc-ms aroma patterns. *Journal of Food Computation and Analysis*, 14:513–522, 2001.

[33] B.K. Lavine, L. Morel, R.K. Vander Meer, R.W. Gunderson, J.H Han, A. Bonanno, and A. Stine. Pattern recognition studies in chemical communication: Nestmate recognition in camponotus floridanus. *Chemometrics and Intelligent Laboratory Systems*, 9:107–114, 1990.

[34] B.K. Lavine, C. Davidson, R.K. Vander Meer, S. Lahav, V. Soroker, and A. Hefetz. Genetic algorithms for deciphering the complex chemosensory code of social insects. *Chemometrics and Intelligent Laboratory Instrumentation*, 66:51–62, 2003.

[35] R.E. Morris, M.H. Hammond, R.E. Shaffer, W.P. Gardner, and S.L. Rose-Pehrsson. The application of chemometric methods to correlate fuel performance with composition from gas chromatography. *Energy and Fuels*, 18:485–489, 2004.

[36] J.A. van Leeuwen, R.J. Jonker, and R. Gill. Octane number prediction based on gas chromatographic analysis with non-linear regression techniques. *Chemometrics and Intelligent Laboratory Systems*, 25:325-340, 1994.

[37] G. Malmquist and R. Danielsson. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *Journal of Chromatography A*, 687:71–88, 1994.

[38] N-P.V. Nielssen, J.M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805:17–35, 1998.

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

[40] P.H.C. Eilers. Parametric time warping. *Analytical Chemistry*, 76:404–411, 2004.

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

[42] J.O. Ramsey, J. ten Berge, and G.P.H. Styan. Matrix correlation. *Psychometrica*, 49:403–423, 1984.

[43] W.J. Krzanowksi. *Principles of Multivariate Analysis, A Users Perspective*. Oxford Science Publications, Oxford, 1998.

[44] E.F. Hilder, F. Svec, and J.H.J. Frechet. Polymeric monolithic stationary phases for capillary electrochromatography. *Electrophoresis*, 23:3924–3953, 2002.

[45] D. Figeys and D. Pinto. Lab-on-a-chip: A revolution in biological and medical sciences. *Analytical Chemistry*, 72:330A–335A, 2000.

[46] L. Tolley, J.W. Jorgenson, and M.A. Moseley. Very high pressure gradient lc/ms/ms. *Analytical Chemistry*, 73:2985–2991, 2001.

[47] L.A. Holland and J.W. Jorgenson. Separation of nanoliter samples of biological amines by a comprehensive two-dimensional microcolumn liquid chromatography system. *Analytical Chemistry*, 67:3275–3283, 1995.

[48] A. van der Horst and P.J. Schoenmakers. Comprehensive two-dimensional liquid chromatography of polymers. *Journal of Chromatography A*, 1000:693–709, 2003.

[49] J.B. Phillips and J. Xu. Environmental applications of comprehensive two-dimensional gas chromatography. *Organohalogen Compounds*, 31:199–, 1997.

[50] P. Korytar, H.-G. Janssen, E. Matisova, and U.A.Th. Brinkman. Practical fast gas chromatography: Methods, instrumentation and applications. *Trends in Analytical Chemistry*, 21:558–572, 2002.

[51] C.A. Cramers, H.-G. Janssen, M.M. van Deursen, and P.A. Leclerq. High-speed gas chromatography: An overview of various concepts. *Journal of Chromatography A*, 856:315–329, 1999.

[52] E. Matisova and M. Domotorova. Fast gas chromatography and its use in trace analysis. *Journal of Chromatography A*, 1000:199–221, 2003.

[53] R. Hoogerbrugge, S.J. Willig, and P.G. Kistemaker. Discriminant analysis by double stage principal component analysis. *Analytical Chemistry*, 55:1710–1712, 1983.

[54] J.T. Scanlon and D.E. Willis. Calculation of flame ionization detector relative response factors using the effective carbon number concept. *Journal of Chromatographic Science*, 23:333–340, 1985.

[55] A.C. Lewis, K.D. Bartle, and A.L. Lee. A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography. *Analytical Chemistry*, 73:1330–1335, 2001.

[56] A.C. van Asten. The importance of gc and gc-ms in perfume analysis. *Trends in Analytical Chemistry*, 21:698–708, 2002.

[57] J. Dalluge, J.J. Vreuls, and U.A.Th. Brinkman. Optimization and characterization of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection. *Journal of Separation Science*, 25:201–214, 2002.

[58] G.S. Frysinger and R.B. Gaines. Prediction of comprehensive two-dimensional gas chromatographic separations. a theoretical and practical exercise. *Journal of Separation Science*, 24:87–96, 2001.

[59] R. Shellie, L. Mondello, P.J. Marriott, and G. Dugo. Characterisation of lavender essential oils by using gas chromatography-mass spectrometry with correlation of linear retention indices and comparison with comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, 970:225–234, 2002.

[60] P.J. Marriott, R. Shellie, R.C.Y. Ong, and P. Morrison. High resolution essential oil analysis by using comprehensive gas chromatographic methodology. *Flavour and Fragrance Journal*, 15:225–239, 2000.

[61] A.J. Kueh, P.J. Marriott, P.M. Wynne, and J.H. Vine. Application of comprehensive two-dimensional gas chromatography to drugs analysis in doping control. *Journal of Chromatography A*, 1000:109–124, 2003.

[62] M. Adachour, J. Beens, R.J.J Vreuls, A.M. Batenburg, E.A.E. Rosing, and U.A.Th. Brinkman. Application of solid-phase micro-extraction and comprehensive two-dimensional gas chromatography for flavour analysis. *Chromatographia*, 55:361–367, 2002.

[63] J. Dalluge, M. van Rijn, J. Beens, J.J. Vreuls, and U.A.Th. Brinkman. Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection applied to the determination of pesticides in food extracts. *Journal of Chromatography A*, 965:207–217, 2002.

[64] J. Blomberg, J. Beens, P.J. Schoenmakers, and R. Tijssen. Comprehensive two-dimensional gas chromatography and its applicability to the characterisation of complex (petrochemical) mixtures. *Journal of High Resolution Chromatography*, 20:539–544, 1997.

[65] J. Beens, R. Tijssen, and J. Blomberg. Prediction of comprehensive two-dimensional gas chromatographic separations: A theoretical and practical exercise. *Journal of Chromatography A*, 822:233–251, 1998.

[66] J.C. Giddings. Sample dimensionality: A predictor of order-disorder in component peak distribution in multidimensional separation. *Journal of Chromatography A*, 703:3–15, 1995.

[67] J. Blomberg, P.J. Schoenmakers, and U.A.Th. Brinkman. Gas chromatographic methods for oil analyis. *Journal of Chromatography A*, 972:137–173, 2002.

[68] L. Mondello, A. Casilli, P.Q. Tranchida, P. Dugo, and G. Dugo. Detailed analysis and group-type separation of natural fats and oils using comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, 1019:187–196, 2003.

[69] H.J. de Geus, I. Aidos, J. de Boer, J.B. Luten, and U.A.Th. Brinkman. Characterisation of fatty acids in biological oil samples using comprehensive multidimensional gas chromatography. *Journal of Chromatography A*, 1019:95–103, 2003.

[70] M. Harju and P. Haglund. Comprehensive two-dimensional gas chromatography of atropisomeric polychlorinated biphenyls, combining a narrow bore beta-cyclodextrin column and a liquid crystal column. *Journal of Microcolumn Separations*, 13(7):300–305, 2001.

[71] E.J. Hayduk, L.H. Choe, and K.H. Lee. Proteomic tools in discovery-driven science. *Current Science*, 83(7):840–844, 2002.

[72] E. Marengo, E. Robotti, P.G. Righetti, and F. Antonucci. New approach based on fuzzy logic and principal component analysis for the classification of two-dimensional maps in health and disease. *Journal of Chromatography A*, 1004:13–, 2003.

[73] B.K. Lavine, A. Vesanen, Brzozowski, and H.T. Mayfield. Authentication of fuel standards using gas chromatography combined with pattern recognition techniques. *Analytical Letters*, 34(2):281–294, 2001.

[74] O. Fiehn, J. Kopka, P. Dormann, T. Altman, R.N. Trethewet, and L. Willmintzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18:1157–1161, 2000.

[75] A.E. Sinha, C.G. Fraga, B.J. Prazen, and R.E. Synovec. Trilinear chemometric analysis of two-dimensional comprehensive gas chromatography coupled to time-of-flight mass spectrometry data. *Journal of Chromatography A*, 1027:269–277, 2004.

[76] K.J. Johnson and R.E. Synovec. Pattern recognition of jet fuels: Comprehensive two-dimensional gas chromatography with anova-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 60:225–237, 2002.

[77] G.S. Frysinger and R.B. Gaines. Determination of oxygenates in gasoline by comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 23:197–201, 2000.

[78] Z. Liu, S.R. Sirmanne, D.G. Patterson, L.L. Needham, and J.B. Phillips. Comprehensive two-dimensional gas chromatography for the fast separation and determienation of pesticides extracted from human serum. *Analytical Chemistry*, 666:3086–3092, 1994.

[79] J. Beens, H. Boelens, R. Tijssen, and J. Blomberg. Quantitative aspects of comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 21:47–54, 1998.

[80] G.S. Frysinger, R.B. Gaines, and E.B. Ledford. Quantitative determination of btex and total aromatic compounds in gasoline by comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 22:195–200, 1999.

[81] T. Hyotylainen, M. Kallio, K. Hartonen, M. Jussila, S. Palonen, and M.J. Riekkola. Modulator design for comprehensive two-dimensional gas chromatography: Quantitative analysis of polyaromatic hydrocarbons and polychlorinated biphenyls. *Analytical Chemistry*, 74:4441–4446, 2002.

[82] C.G. Fraga, B.J. Prazen, and R.E. Synovec. Comprehensive two-dimensional gas chromatography and chemometrics for the high-speed quantitative analysis of aromatic isomers in a jet fuel using the standard addition method and an objective retention time alignment algorithm. *Analytical Chemistry*, 72:4154–4162, 2000.

[83] C.G. Fraga, C.A. Bruckner, and R.E. Synovec. Increasing the number of analyzable peaks in comprehensive two-dimensional separations through chemometrics. *Analytical Chemistry*, 73:675–683, 2001.

[84] R. Shellie, L.L. Xie, and P.J. Marriott. Retention time reproducibility in comprehensive two-dimensional gas chromatography using cryogenic modulation: An intralaboratory study. *Journal of Chromatography A*, 968:161–170, 2002.

[85] C.G. Fraga, B.J. Prazen, and R.E. Synovec. Enhancing the limit of detection for comprehensive two-dimensional gas chromatography data using bilinear chemometric analysis. *Journal of High Resolution Chromatography*, 23:215–224, 2000.

[86] H.A.L. Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14:105–122, 2000.

[87] P.J. Schoenmakers, P.J. Marriott, and J. Beens. Nomenclature and conventions in comprehensive multidimensional chromatography. *LCGC Europe*, 16:335–, 2002.

[88] S. Macho and M.S. Larrechi. Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *Trends in Analytical Chemistry*, 21:799–806, 2002.

[89] H.G. Law, C.W. Snyder, J. Hattie, and R.P. McDonald. *Research Methods for Multimode Analysis*. John Wiley and Sons, New York, 1984.

[90] P.D. Wentzell, S.S. Nair, and R.D. Guy. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Analytical Chemistry*, 73:1408–1415, 2001.

[91] R. Bro. Parafac: Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38:149–171, 1997.

[92] V. Pravdova, C. Boucon, S. de Jong, B. Walczak, and B.L. Massart. Three-way principal component analysis applied to food analysis : An example. *Analytica Chimica Acta*, 462:133–148, 2002.

[93] A.K. Smilde, P.H. van der Graaf, D.A. Doornbos, T. Steerneman, and A. Sleurink. Multivariate calibration of reversed-phase chromatographic systems: Some designs based on three-way data analysis. *Analytica Chimica Acta*, 235:41–51, 1990.

[94] A.K. Smilde and D.A. Doornbos. Three-way methods for the calibration of chromatographic systems: Comparing parafac and three-way pls. *Journal of Chemometrics*, 5:345–360, 1991.

[95] R. Bro. Multi-way calibration. multi-linear pls. *Journal of Chemometrics*, 10:47–62, 1996.

[96] A.K. Smilde and H.A.L. Kiers. Multiway covariates regression models. *Journal of Chemometrics*, 13:31–48, 1999.

[97] J.D. Carrol and J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckardt-young" decomposition. *Psychometrica*, 35:283–319, 1970.

[98] R.A. Harshman. Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis. *Working Papers in Phonetics*, 16:1–84, 1970.

[99] H.A.L. Kiers, J.F. ten Berge, and R. Bro. Part 1: A direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13:275–294, 1999.

[100] Andersson C.A. Bro, R. and H.A.L. Kiers. Part 2: Modeling chromatographic data with retention time shifts. *Journal of Chemometrics*, 13:295–309, 1999.

[101] B.W. Wise, N.B. Gallaher, and E.B. Martin. Application of parafac2 to fault detection and diagnosis in semiconductor etch. *Journal of Chemometrics*, 15:285–298, 2001.

[102] C.G. Zampronio, S.P. Gurden, L.A. Moraes, M.N. Eberlin, A.K. Smilde, and R.J. Poppi. Direct sampling tandem mass-spectrometry and multiway calibration for isomer quantitation. *Analyst*, 127:1054–1060, 2002.

[103] J.B. Phillips, R.B. Gaines, J. Blomberg, F.W.M. van der Wielen, J.M. Dimandja, V. Green, J. Granger, D. Patterson, L. Racovalis, H.J. de Geus, P. Haglund, J. Lipsky, V. Sinha, and E.D. Ledford. A robust thermal modulator for comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 22(1):3–10, 1999.

[104] U.s. geological survey, woodshole ma 02543.

[105] C.A. Andersson and R. Bro. The n-way toolbox for matlab. *Chemometrics and Intelligent Laboratory Systems*, 50:1–4, 2000.

[106] D. Bylund, R. Danielsson, and K.E. Markides. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatography - mass spectrometry data. *Journal of Chromatography A*, 961:237–244, 2002.

[107] J. Beens, J. Blomberg, and P.J. Schoenmakers. Proper tuning of comprehensive two-dimensional gas chromatography to optimize the separation of complex oil fractions. *Journal of High Resolution Chromatography*, 23:182–188, 2000.

[108] R. Shellie and P.J. Marriott. Opportunities for ultra-high resolution analysis of essential oils using comprehensive two-dimensional gas chromatography. *Flavour and Fragrance Journal*, 18:179–191, 2003.

[109] J.F. Hamilton and A.C. Lewis. Monoaromatic complexity in gasoline and urban air using comprehensive two-dimensional gas chromatography and gas-chromatography coupled to time-of-flight mass spectrometry. *Atmospheric Environment*, 37:589–602, 2003.

[110] S.E. Reichenbach, M. Ni, D. Zhang, and E.B. Ledford. Image background removal in comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, 985:47–56, 2003.

[111] S.E. Reichenbach, M. Ni, V. Kottapalli, and A. Visvanathan. Information technologies for comprehensive two-dimensional gas chromatography. *Chemometrics and Intelligent Laboratory Systems*, 71:107–120, 2004.

[112] G. Malmquist. Multivariate evaluation of peptide mapping using the entire chromatographic profile. *Journal of Chromatography A*, 687:89–100, 1994.

[113] L.M. Blumberg and M.S. Klee. Method translation and retention time locking in partition gc. *Analytical Chemistry*, 70:3828–3839, 1998.

[114] J. Dalluge, L.L.P. van Stee, X. Xu, J. Williams, J. Beens, J.J. Vreuls, and U.A.Th. Brinkman. Unravelling the composition of very complex samples by comprehensive gas chromatography coupled to time-of-flight mass spectrometry. cigarette smoke. *Journal of Chromatography A*, 974:169–184, 2002.

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

[116] A. Goshtasby. Image registration by local approximation methods. *Image and Vision Computing*, 6:255–261, 1988.

[117] Image processing toolbox, the mathworks.

[118] R.C.Y. Ong and P.J. Marriott. A review of basic concepts in comprehensive two-dimensional gas chromatography. *Journal of Chromatographic Science*, 13:276–291, 2002.

[119] G. Grob, K. Grob, and K. Grob. Comprehensive, standardized quality test for glass capillary columns. *Journal of Chromatography A*, 156:1–20, 1978.

[120] C.G. Harrigan and R. Goodacre. *Metabolic Profiling: Its Role to Biomarker Discovery and Gene Function Analysis.* Kluwer Academic Publishing, Boston, 2003.

[121] M. Daszykowski, B. Walczak, and D.L. Massart. Projection methods in hemistry. *Chemometrics and Intelligent Laboratory Systems*, 65:97–112, 2003.

[122] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Seyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A.* Elsevier, Amsterdam, 1997.

[123] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Seyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part B.* Elsevier, Amsterdam, 1998.

[124] K.A. Anderson and B.W. Smith. Chemical profiling to differentiate geographic growing origin of coffee. *Journal of Agricultural Food Chemistry*, 50:2068–2075, 2002.

[125] A.J. Charlton, W.H.H. Farrington, and P. Brereton. Application of (1)h nmr and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee. *Journal of Agricultural Food Chemistry*, 50:3098–3103, 2002.

[126] J.T.W.E. Vogels, A.C. Tas, F. van den Berg, and J. van der Greef. A new method for classification of wines based on proton and carbon-13 nmr spectroscopy in combination with pattern recognition techniques. *Journal of Chemometrics and Intelligent Laboratory Systems*, 21:249–258, 1993.

[127] M.A. Brescia, V. Caldarola, A. de Giglio, D. Benedetti, F.P. Fanizzi, and A. Sacco. Characterization of the geographical origin of italian red wines based on traditional and nuclear magnetic resonance spectrometric determinations. *Analytica Chimica Acta*, 458:177–186, 2002.

[128] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley and Sons, New-York, 1992.

[129] M. Barker and W. Raynes. Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173, 2003.

[130] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.