New Chemometric Approaches to Non-targeted GCMS Fingerprinting Analysis of Wine Volatiles

Jochen Vestner

New Chemometric Approaches to Non-targeted GC-MS Fingerprinting Analysis of Wine Volatiles

par

Jochen Vestner

Thèse en cotutelle entre

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

et

HOCHSCHULE GEISENHEIM UNIVERSITY

pour obtenir le grade de

DOCTEUR

Mention: Sciences, Technologie, Santé
Option: Œnologie

Présentée et soutenue publiquement
Le 13 septembre 2016

Membres du Jury:

P. Guedes de Pinho	Directeur de Recherche, Universidade do Porto	Présidente
A.C.S. Ferreira	Professeur, Universidade Católica Portuguesa & Stellenbosch University	Rapporteur
F. Mattivi	Professeur, Fondazione Edmund Mach (Italie)	Rapporteur
G. de Revel	Professeur, Université de Bordeaux	Directeur de thèse
D. Rauhut	Professeure, Hochschule Geisenheim University	Directrice de thèse

Titre : Nouvelles approaches par empreinte chromatographique non ciblées des composés volatiles du vin

Résumé :

Contrairement à l'analyse ciblée des composés volatils du vin par chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS), les approches par GC-MS non ciblées prennent en compte les composés connus et inconnus. Ces méthodes sont plus rapides et fournissent une représentation plus complète de la composition de l'échantillon. Bien que plusieurs approches non-ciblées aient été développées, il y a encore une forte demande d'outils automatisés pour le traitement des données, en particulier pour les données multidimensionnelles complexes telles que celles de multiples chromatogrammes GC-MS.

Ce travail visait à développer deux nouvelles approches chimiométriques pour l'analyse des données GC-MS non ciblées. Ces approches prennent en considération les décalages de temps de rétention entre les échantillons et rendent inutile l'intégration des pics. Elles ont été testées avec un jeu de données GC-MS simulées et un jeu de données GC-MS réelles d'échantillons de vin.

De plus, l'une des deux approches GC-MS non ciblée a été combinée à la technique d'analyse sensorielle rapide de "projective mapping". Cette méthodologie a été utilisée pour étudier l'impact de la fermentation malolactique sur des vins issus du cépage Pinotage ainsi que l'effet de l'âge de la vigne, de la turbidité du moût et de la souche de levure sur l'arôme de vins de Riesling expérimentaux.

Mots clés : composés volatiles du vin, GC-MS, analyse non ciblée

---

Title : New Chemometric Approaches to Non-targeted GC-MS Fingerprinting Analysis of Wine Volatiles

Abstract :

In contrast to targeted gas chromatography mass spectrometry (GC-MS) analysis of wine volatiles, non-targeted GC-MS approaches take information of known and unknown compounds into account, are faster, inherently more comprehensive and give a more holistic representation of the sample composition. Although several non-targeted approaches have been developed, there is still a great demand for automated data processing tools, especially for complex multi-way data such as

chromatographic data obtained from multichannel detectors (e.g. GC-MS chromatograms of multiple samples).

This work therefore aimed at the development of data processing procedures for non-targeted GC-MS analysis of volatile wine compounds. The two developed approaches use basic matrix manipulation of segmented GC-MS chromatograms and PCA or PARAFAC multi-way modelling. The approaches take retention time shifts between samples into account and avoid peak integration. A demonstration of the new fingerprinting approaches is presented using an artificial GC-MS data set and an experimental full-scan GC-MS data set obtained for a set of experimental wines. Results of the new approaches were also compared to a references method.

Furthermore, the combination of one of the developed GC-MS fingerprinting approaches with the fast sensory screening technique projective mapping was exploited as a powerful approach to simultaneously study the volatile composition and the sensory characteristics of experimental wines. This methodology was used to study the impact of different malolactic fermentation scenarios on two different Pinotage wine styles and for a full factorial investigation of the impact of grape vine age, must turbidity and yeast strain on the aroma of Riesling experimental wines.

Keywords : wine volatiles, GC-MS, non-targeted analysis

Unité de recherche :

Unite de Recherche Œnologie, EA 4577, USC 1366 INRA

ISVV

210, chemin de Leysotte

CS 50008

33882 Villenave d'Ornon, France

[Intitulé, n° de l’unité, et adresse de l’unité de recherche]

The answer to the Ultimate Question of Life, the Universe, and Everything:

'*Forty-two*, said Deep Thought, with infinite majesty and calm.'

The Hitchhiker's Guide to the Galaxy. Douglas Adams (1979).

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following people and organizations for their contribution throughout this study:

My supervisors, Prof. Dr. Doris Rauhut and Prof. Dr. Gilles de Revel, for their supervision and for providing me the rare opportunity of a co-supervised thesis between two European universities. I would like to thank them particularly for all the effort and time they have invested into the realization of this cotutelle.

Prof. Dr. Gilles de Revel for taking me into his amazing team and for giving me a very focused view on oenology research.

Prof. Dr. Doris Rauhut for her support for many years from my undergraduate studies until my PhD. For giving me the freedom for creativity and an independent realization of this thesis.

The members of the doctoral committee, Dr. Paula Guedes de Pinho, Prof. Dr. António César da Silva Ferreira and Dr. Fulvio Mattivi for the evaluation of my thesis and for providing a very valuable scientific discussion during the defence.

Prof. Dr. André de Villiers, who has been a great mentor since my M.Sc., for his helpful advices and ongoing collaborations. Martha, Chandré, Andreas and the rest of the group for a great time Stellenbosch.

Dr. Armin Schüttler for all his help with Chapter 5, a good time sharing an office and several 'Kellerbesprechungen'.

The research group at the ISSV in Bordeaux for a warmly welcome, all the patience with my moderate french, several 'Aperos' and loads of fun.

The research group in Geisenheim for help in the laboratory. Particularly Christopher Geus is thanked for the help with the work in Chapter 5 and supply with franconian beer.

Prof. Dr. Maret du Toit and her team at the IWBT of the Stellenbosch University

for valuable scientific discussions and supplying the wines in Chapter 4.

Dr. Claus Patz and Matthias Friedel for discussions on chemometrics and MATLAB.

Dr. Sibylle Kreiger-Weber and team from Lallemand for supplying the wines in Chapter 3 and valuable discussions on malolactic fermentation.

Prof. Dr. Rasmus Bro for giving me the possibility of attending two chemometric courses at the Department of Food Science at the University of Copenhagen, amazing discussions chemometrics and inspirations.

All fellows at the ISSV for support during the sensory studies.

All fellows at the department of grapevine breeding at the Hochschule Geisenheim University, particularly Roger Grundel, for support with experimental winemaking.

Julius Witte and Kimmo Sirén for help and discussion on matrix algebra.

The Department of general and organic viticulture at the Hochschule Geisenheim University for supplying grapes and assistance with the winemaking in Chapter 5.

Prof. Dr. Deirdre Cabooter for the stay in her laboratory and a good time in Leuven.

Prof. Dr. Tadeusz Górecki for great discussions on gas chromatography.

Everybody who was involved in administrational work related to this cotutelle and the Oenoviti International Network.

Initiative d'excellence de l'Universit de Bordeaux (IDEX Bordeaux) and the Hochschlule Geisenheim for funding.

My family and Kathrin for their support, patience and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF ABBREVIATIONS

ALS Alternating Least Squares

AMDIS Automated Mass Spectral Deconvolution and Identification System

CA Correspondence Analysis

CCSWA Common Components and Specific Weights Analysis

EFA Evolving Factor Analysis

FID Flame Ionisation Detector

GC Gas Chromatography

GC-MS Gas Chromatography Mass Spectrometry

GPA General Procrustes Analysis

HCA Hierarchical Cluster Analysis

HS-SPME-GC-MS Headspace Solid Phase Microextraction Gas Chromatography Mass Spectrometry

HS-SPME-GC×GC-TOFMS Headspace Solid Phase Microextraction Comprehensive Two-dimensional Gas Chromatography Time-of-flight Mass Spectrometry

LAB Lactic Acid Bacteria

LC-MS Liquid Chromatography Mass Spectrometry

LLE Liquid-Liquid Extraction

MCR Multivariate Curve Resolution

MDS Multidimensional Scaling

MFA Multiple Factor Analysis

MLF Malolactic Fermentation

N-PLS Multi-way Partial Least Squares

NMR Nuclear Magnetic Resonance

OPLS-DA Orthogonal Partial Least Squares Discriminant Analysis

PARAFAC Parallel Factor Analysis

PARAFAC2 Parallel Factor Analysis 2

PC Principal Component

PCA Principal Component Analysis

PLS Partial Least Squares

PLS-DA Partial Least Squares Discriminant Analysis

QDA Quantitative Descriptive Analysis

SIMCA Soft Independent Modeling of Class Analogy

SPE Solid Phase Extraction

SPME Solid Phase Microextraction

SSCP Sums of Squares and Cross Products

SVD Singular Value Decomposition

TIC Total Ion Chromatogram

SCIENTIFIC COMMUNICATION

Publications related to the subject of this thesis

2016 **Jochen Vestner**, Sibylle Krieger-Weber, Gilles de Revel, Doris Rauhut, André de Villiers. Toward Automated Chromatographic Fingerprinting: A Non-Alignment Approach to Gas Chromatrography Mass Spectrometry Data. *Analytica Chimica Acta*, 911, 42-58

Other publications

2015 Chandré M. Willemse, Maria A. Stander, **Jochen Vestner**, André de Villiers. Comprehensive two-dimensional HILIC×RP-LC-UV-MS analysis of anthocyanins and derived pigments in red wine. *Analytical Chemistry*, 87 (24), 12006-12015

2013 Kathithileni M. Kalili; **Jochen Vestner**; Maria A. Stander; André de Villiers. Toward Unraveling Grape Tannin Composition: Application of Online Hydrophilic Interaction Chromatography × Reversed-Phase Liquid ChromatographyTime-of-Flight Mass Spectrometry for Grape Seed Analysis. *Analytical Chemistry*, 85 (19), 9107-9115

2011 **Jochen Vestner**; Sulette Malherbe; Maret Du Toit; Hélène H. Nieuwoudt; Ahmed Mostafa; Tadeusz Górecki; Andreas G. J. Tredoux; André de Villiers. Investigation of the volatile composition of Pinotage wines fermented with different malolactic starter cultures using comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC-TOF-MS). *Journal of Agricultural and Food Chemistry*, 59, 12732-12744

2010 Andrew C. Clark; **Jochen Vestner**; Célia Barril; Chantal Maury; Paul D. Prenzler; Geoffrey R. Scollary. The influence of stereochemistry of antioxidants and flavanols on oxidation processes in a model wine system: ascorbic acid, erythorbic acid, (+)-catechin and (-)-epicatechin. *Journal of Agricultural and Food Chemistry*, 58, 1004-11

2010 **Jochen Vestner**; Stefanie Fritsch; Doris Rauhut. Development of a microwave assisted extraction method for the analysis of 2,4,6-trichloroanisole in cork stoppers by SIDA-SBSE-GC-MS. *Analytica Chimica Acta*, 660, 76-80

Oral and poster communications

2015 9th In Vino Analytica Scientia Symposium (IVAS2015), Trentino, Italy (Oral)

2015 10th International Symposium of Enology (OENO2015), Bordeaux, France (Oral)

2014 36th Congress of the South African Society for Enology and Viticulture (SASEV), Stellenbosch, South Africa (Poster)

2013 8th In Vino Analytica Scientia Symposium, Reims, France (Oral)

2012 34th Congress of the South African Society for Enology and Viticulture (SASEV), Stellenbosch, South Africa (Poster)

2012 Hyphenated Techniques for Chromatography (HTC-12) & Hyphenated Techniques for Sample Preparation (HTSP-2), Bruges, Belgium (Poster)

2012 36th International Symposium on Capillary Chromatography (ISCC) and 9th GCxGC Symposium, Riva, Italy (Poster)

2010 Analitika, Stellenbosch, South Africa (Poster)

2010 32th Congress of the South African Society for Enology and Viticulture (SASEV), Stellenbosch, South Africa (Poster)

ABSTRACT

New Chemometric Approaches to Non-targeted GC-MS Fingerprinting Analysis of Wine Volatiles

by

Jochen Vestner

In contrast to targeted gas chromatography mass spectrometry (GC-MS) analysis of wine volatiles, non-targeted GC-MS approaches take information of known and unknown compounds into account, are faster, inherently more comprehensive and give a more holistic representation of the sample composition. Although several non-targeted approaches have been developed, there is still a great demand for automated data processing tools, especially for complex multi-way data such as chromatographic data obtained from multichannel detectors (e.g. GC-MS chromatograms of multiple samples).

This work therefore aimed at the development of data processing procedures for non-targeted GC-MS analysis of volatile wine compounds. The two developed approaches use basic matrix manipulation of segmented GC-MS chromatograms and PCA or PARAFAC multi-way modelling. The approaches take retention time shifts between samples into account and avoid peak integration. A demonstration of the new fingerprinting approaches is presented using an artificial GC-MS data set and an experimental full-scan GC-MS data set obtained for a set of experimental wines.

Results of the new approaches were also compared to a references method.

Furthermore, the combination of one of the developed GC-MS fingerprinting approaches with the fast sensory screening technique projective mapping was exploited as a powerful approach to simultaneously study the volatile composition and the sensory characteristics of experimental wines. This methodology was used to study the impact of different malolactic fermentation scenarios on two different Pinotage wine styles and for a full factorial investigation of the impact of grape vine age, must turbidity and yeast strain on the aroma of Riesling experimental wines.

RÉSUMÉ

Nouvelles approaches par empreinte chromatographique non ciblées des composés volatiles du vin

par

Jochen Vestner

Contrairement à l'analyse ciblée des composés volatils du vin par chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS), les approches par GC-MS non ciblées prennent en compte les composés connus et inconnus. Ces méthodes sont plus rapides et fournissent une représentation plus compléte de la composition de l'échantillon. Bien que plusieurs approches non-ciblées aient été développées, il y a encore une forte demande d'outils automatisés pour le traitement des données, en particulier pour les données multidimensionnelles complexes telles que celles de multiples chromatogrammes GC-MS.

Ce travail visait à développer deux nouvelles approches chimiométriques pour l'analyse des données GC-MS non ciblées. Ces approches prennent en considération les décalages de temps de rétention entre les échantillons et rendent inutile l'intégration des pics. Elles ont été testées avec un jeu de données GC-MS simulées et un jeu de données GC-MS réelles d'échantillons de vin.

De plus, l'une des deux approches GC-MS non ciblée a été combinée à la technique d'analyse sensorielle rapide de "projective mapping". Cette méthodologie a

été utilisée pour étudier l'impact de la fermentation malolactique sur des vins issus du cépage Pinotage ainsi que l'effet de l'âge de la vigne, de la turbidité du moût et de la souche de levure sur l'arôme de vins de Riesling expérimentaux.

CHAPTER I

Introduction

1.1 General introduction

Wine has been an essential part of the sophisticated way of life in many cultures for thousands of years. From the ancient Egyptians to European cultures nowadays, wine consumption and production has, however, changed significantly. Compared to the ancient Retsina wines, which were also used as medicine, today's high quality wines are made to meet sensory expectations of the modern consumer. In other words, winemakers want to meet consumer preferences for distinct wine styles. As a consequence, the most important quality driver of modern wine is its aroma. The modulation of wine aroma presupposes a vast understanding of the volatile composition of wine and the impact of viticultural and oenological influencing factors.

The analytical method of choice for the analysis of volatiles is gas chromatography. Since the introduction of commercial gas chromatography instruments in the late 1950s targeted methods for several wine volatiles have been developed, which always presuppose an *a priori* known and identified set of compound. Until today, targeted methods, which have the major advantage of accurate quantifications, are mainly used in wine aroma research. This advantage however comes along with the disadvantage of time consuming calibration procedures and the fact that information about differences among samples can only be obtained for a limited number of compounds. The steep

rise of metabolomics in the last two decades also inspired wine scientists to use non-targeted approaches for the analysis of wine volatiles. Non-targeted analysis aims to gather qualitative and (semi-)quantitative information on as many compounds as possible in the analysed samples in a short period of time, and thus to provide the researcher with a more holistic view of the composition of samples. Non-targeted strategies are therefore more comprehensive and can be hypothesis generating, as semi-quantitative information on a wide range of different compounds is obtained. Considering the complexity of the wine matrix which includes hundreds of volatile compounds, non-targeted approaches can be useful to shed new light into the research of wine aroma.

Recent advances in the development of analytical instrumentation enable fast, accurate and cost effective analyses of a large number of samples in numerous domains of analytical chemistry. These improvements in technology made non-targeted screening and fingerprinting analyses of large sample sets possible in the first place, but also lead to a vast increase of more complex data which has to be processed and analysed. The conventional way of addressing these big datasets includes chromatographic preprocessing such as retention time alignment, feature selection (e.g. peak picking) and multivariate modelling of the final peak table. This conventional strategy is also implemented in the available software packages for data analysis of non-targeted chromatographic analysis. Retention time alignment is sometimes difficult to apply and prone to errors (wrong assignment of peaks), while applying feature selection information can be missed, as all peaks missing a certain criteria are not taken into account in further multivariate analysis. These disadvantages of the conventional strategy for non-targeted data analysis indicate the necessity of novel data analysis approaches.

1.2 Objectives of this study

The principle objective of this dissertation was the development of a new data analysis approach for non-targeted fingerprinting GC-MS analysis of wine volatiles to overcome drawbacks of conventional methods concerning retention time alignment and feature selection. The alignment issue was solved by segmenting chromatograms and their transformation using linear algebra. By transforming segments of the two-dimensional chromatographic signal of each sample into Sums of Squares and Cross Products (SSCP) matrices, a measure for variations of the mass channels and covariations among the mass channels were obtained for each segment. The sums of squares and cross products of the mass channels are not afected by the location of peaks in the segments. Peak shifts among samples do not therefore influence these measures of variation within a mass channel and covariation among mass channels. Based on this transformation, two approaches were developed. Approach one includes further rearrangement of the matrices resulting in a three-way array which can be directly decomposed using the multi-way method Parallel Factor Analysis (PARAFAC). In approach two the SSCP matrices are decomposed in a singular value decomposition (SVD) for each segment and sample and only the first singular values are kept for further principal component analysis (PCA). Both approaches avoid peak alignment and feature selection such as peak integration and were tested on an artificial and a real GC-MS data set. The PARAFAC model in approach one is more difficult to model, but reveals more information on systematic differences among samples and can be used with supervised as well as with unsupervised preprocessing. For approach two supervised preprocessing is inevitable. This approach can therefore only be used when samples can be categorized in classes.

GC-MS fingerprintings of wine volatiles provide important analytical data. It is however not possible to draw conclusions regarding the sensory properties of wines from analytical data alone. The linkage of analytical and sensory data is in this

regard a important necessity in wine aroma research. A further aim of this thesis was therefore the integration of results from the developed GC-MS fingerprinting approach with rapid sensory screenings such as partial projective mapping to obtain a more holistic view on the aroma of wines. The combination of these two techniques provides an fast and efficient tool for multi-parametric aroma studies of experimental wines. Two relevant topics of interest in wine research have been addressed. The first application was the investigation of the impact of different malolactic fermentation (MLF) starter cultures and inoculation scenarios on the aroma expression of two different Pinotage styles. The second application comprised experimental wine making in full factorial design to study the effects of grapevine age, turbidity and yeast starter culture on the aroma of Riesling wines and how these factors influence each other.

1.3 Thesis outline

Chapter 2 - Literature review

This chapter gives an introduction on chromatographic data analysis with an particular focus on conventional and alternative methods for non-targeted data analysis. Moreover, an overview of volatile wine constituents contributing to the aroma of wine and their analysis using gas chromatography is provided. Finally, a short introduction on rapid sensory profiling techniques is given.

Chapter 3 - Development of new approaches to non-targeted GC-MS data analysis

Chapter 3 constitutes the major part of the thesis and deals with the development of new approaches for non-targeted GC-MS data analysis which consider retention time shifts and avoid feature selection such as peak picking. After the background considerations on the used strategy are presented, both approaches are tested on an artificial and a real GC-MS data set and validated with an reference method.

Chapter 4 - Application 1: Comparative aroma study on the impact of different malolactic fermentation scenarios on two Pinotage wine styles

The influence of different MLF starter cultures and inoculation modes on the aroma of Pinotage wines is investigated. Results of the differences between the volatile composition of samples obtained from the first approach are combined with the results from fast sensory screening (perceptual mapping) using the multi-block PCA method multiple factor analysis (MFA).

Chapter 5 - Application 2: Full factorial aroma study on the impact of grapevine age, yeast strain and must turbidity on the aroma of Riesling experimental wines

The developed strategy of integrating GC-MS fingerprinting results with those of fast sensory screening from perceptual mapping of wines (Chapter 4) is extended to experimental wine making in full factorial design. Main and interaction effects of the factors grapevine age, yeast strain and must turbidity on the volatile composition and the aroma expression of the Riesling experimental wines were studied.

Chapter 6 - General conclusions

A summary of results and the major findings is given from the development of the non-targeted data analysis approaches to the application of the first approach to the Pinotage and Riesling experimental wines, and data merging of chemical and sensory data.

CHAPTER II

Literature review

From the beginning of commercial Gas Chromatography (GC) in the 1950's until today the principal visual appearance of chromatographic signals has not changed. The chromatographic signal is visualized as a time series where the detector deflect is represented as peaks corresponding to the eluting substances. In the case of multichannel detectors, such as a mass spectrometer multiple scans are captured in sequence. Then as now, chromatographers have to extract qualitative and quantitative information from chromatograms such as the identities or the concentrations of compounds. Data processing methods have, however, significantly changed in the last 60 years from trivial methods such as cutting out peaks and weighting the cut paper (Carroll, 1961) to computer modelling of extremely rich data sets of chromatograms from targeted and non-targeted studies in metabolic research[1] nowadays (Eliasson et al., 2011). And yet, there is still a lack of fast and automated data processing approaches for chromatographic data, in particular for non-targeted analysis.

The intention of this Chapter is not to discuss chromatographic theory, instru-

[1]In metabolic research quantifications (absolute or relative) of one or a few target compounds in a series of biological samples are called *metabolite target analysis*, while the quantitative (absolute or relative) and qualitative multi-component analysis that define or describe metabolic patterns for a group of metabolically or analytically related metabolites is called *metabolic profiling* (Horning and Horning, 1971). *Metabolic fingerprinting* is high throughput screening for sample classification by spectroscopic techniques such as NMR or direct infusion mass spectrometry (DIMS). The term *metabolomics* refers to non-targeted qualitative and quantitative analysis of the complete set of metabolites present in a biological system (Dunn and Ellis, 2005; Fiehn, 2002; Koek et al., 2011).

mentation and optimization of separation, which can be found in more dedicated literature (Sparkman et al., 2011; Hübschmann, 2008), but to give an overview of methods for targeted and non-targeted chromatographic data analysis. Moreover, the application of GC and sensory analysis in wine aroma research is reviewed.

2.1 Conventional targeted chromatographic data analysis

In conventional quantitative targeted analysis chromatographic peaks are usually fully separated and integrated from the beginning to the end of the peak. The peak areas of samples with known concentrations are used to build a calibration curve and peak areas of unknown samples are related to the calibration curve to determine the accurate concentration of a compound. Peak integration and calculations of concentrations are usually done using commercial software provided from the manufacturer of the chromatographic system. The amount of a compound in a sample can be stated as the accurate concentration (e.g. $\mathrm{mg\,L^{-1}}$). The biggest advantage of accurate targeted quantification is the comparability of results among measured sequences of samples, instruments and laboratories. Disadvantages are that the identity of the component has to be known, standards of known purity have to be available and the calibration procedure is usually time consuming. In some cases, when no standard of known concentration is available, compounds can also be calibrated with reference standards. Such a reference standard is usually a structurally similar compound. Concentrations are then expressed as concentrations calculated relative to the reference standard.

2.2 Non-targeted and multivariate chromatographic data analysis

Non-targeted analysis has increasingly gained importance in numerous domains of analytical chemistry such as life science, food science and especially the '-omics' related sciences. In contrast to conventional targeted analysis, non-targeted analysis aims to gather qualitative and quantitative information on as many compounds as possible in the analysed samples in a short period of time, and thus to provide the researcher with a more holistic view of the composition of samples (De Vos et al., 2008). Holistic strategies benefit from the vast amount of information obtained from modern analytical instrumentation. And yet the main challenges associated with non-targeted analysis are data handling and full exploitation of dimensionality of the acquired data. Modern chromatographic instruments such as GC-MS allow automated, reproducible and fast analysis of many samples and are therefore especially suited for non-targeted approaches.

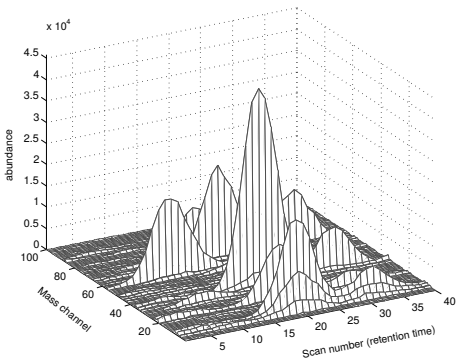
Conventional analysis of non-targeted chromatographic data, such as the common GC-MS metabolomics workflows, generally includes certain steps of data preprocessing such as noise filtering, baseline correction, alignment of peaks, feature selection (e.g. peak detection), identification of peaks, normalization prior to multivariate data analysis and interpretation of the results (Koek et al., 2011). Conventional data analysis approaches and available software packages for non-targeted GC-MS analysis are reviewed in Section 2.2.2. Nevertheless, there is an increasing tendency to use the entire chromatographic profile as a chemical fingerprint containing a unique pattern characteristic for a sample. Benefits and difficulties of fingerprinting approaches are further discussed in Section 2.2.3. But first, an overview on chromatographic data structure is given in Section 2.2.1.

2.2.1 Chromatographic data structure

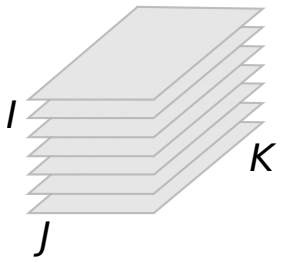
When multi-channel detectors are used, the chromatographic separation and the detector provide different dimensions of data. Different representations of a section of a two-dimensional GC-MS chromatogram are shown in Figure 2.1. A GC-MS chromatogram can be considered as a matrix of dimensions, *scan number* $\times$ *mass channels*. A data set of multiple two dimensional GC-MS chromatograms can consequently be represented as a three-way array (Figure 2.1(b)). Single channel detectors such as the Flame Ionisation Detector (FID) simply produce a time resolved signal similar to the Total Ion Chromatogram (TIC) of a GC-MS chromatogram, which represents the elution time profile of the summed MS dimension (Figure 2.1(c)).

2.2.2 Conventional non-targeted chromatographic data analysis

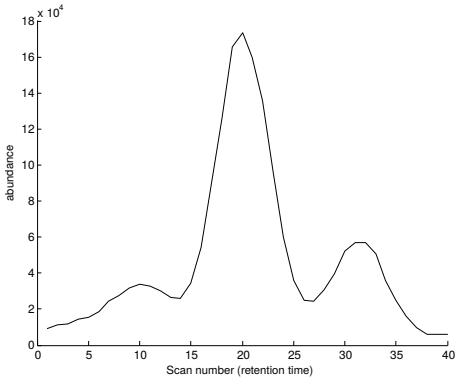
The steep rise of metabolomics during the last two decades is closely linked with the continuous development of modern analytical instrumentation, especially the advances in GC-MS and Liquid Chromatography Mass Spectrometry (LC-MS). The necessity of processing more opulent data from more complex instrumentation leads to the development of new algorithms and software tools for non-targeted metabolomics data. Besides commercial software, many free and open source software packages are available today. The probably best known software packages are XCMS (Smith et al., 2006), its extensions such as metaMS (Wehrens et al., 2014), MZmine (Katajamaa et al., 2006; Pluskal et al., 2010) and MetAlign (Lommen, 2009); many others are listed for instance in Niu et al. (2014); Theodoridis et al. (2012); Castillo et al. (2011). Depending on the chromatographic system and the size of the sample set noise, baseline drift and retention time shifts of peaks among samples are common problems decreasing the quality of chromatograms. Most software for non-targeted chromatographic data analysis address therefore certain preprocessing steps including noise reduction, baseline correction, alignment of peaks, feature selection such as peak



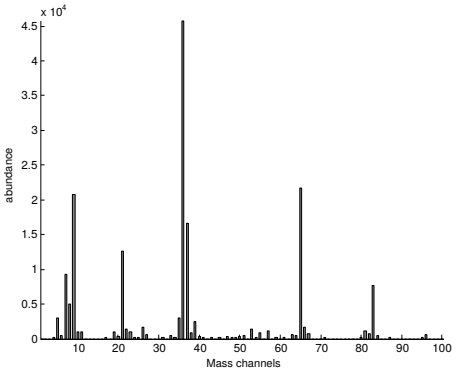
(a) Two dimensional representation, elution profile vs. mass channels



(b) Three-way array: I samples $\times$ J scans (elution profile) $\times$ K mass channels



(c) TIC (summed MS dimension)



(d) Mass spectra of scan number 20

Figure 2.1: Different representations of a two-dimensional GC-MS chromatogram section (peak system) consisting of 40 scans (time points) and 100 mass channels.

detection, identification of peaks, and normalization prior to multivariate modelling of peak area tables. These steps may be performed in a different order, depending on the data analysis strategy. For instance alignment can also be done after peak detection.

A chromatographic signal consists of an analytical signal, baseline and noise. Preprocessing in conventional non-targeted data analysis aims to eliminate irrelevant variations in the chromatogram caused by noise and background interferences in order to properly extract important analyte information and chemical variations. Baseline correction methods adjust baseline drift and reduce low frequency background variations in the chromatogram usually caused by column bleeding, background ionization and low frequency detector variations. The simplest way of addressing baseline distortion is the subtraction of a blank chromatogram from the sample chromatograms. As this simple method is not always applicable, the most commonly used baseline correction is a polynomial least square fitting to simulate a blank chromatogram. Subsequently, the fitted baseline is subtracted from the sample chromatogram. Important for any baseline correction algorithm is to avoid overfitting of the baseline and any elimination and alteration of chemical relevant information. Additionally, factor models can be used to deconvolute baseline and analytical signal in sub-regions of the chromatogram (see Section 2.2.4 for more details).

Noise filtering and smoothing are performed to increase the signal-to-noise ratio by removing high frequency noise from the signal. The most widely used noise reduction technique is the classical Savitzky-Golay method, which fits a least squares polynomial of a given order to a certain window size in the chromatogram (Savitzky and Golay, 1964). Other noise reduction methods are based on wavelet smoothing (Barclay et al., 1997). Wavelet smoothing algorithms transform the chromatogram into the frequency domain, removing the high-frequency noise, and reverting back to the retention time domain with the result of an smoothed chromatogram.

Small retention time variations among chromatographic runs are generally unavoidable due to column ageing, uncontrollable pressure, flow and temperature fluctuations. Retention time shifts are even more severe in LC analysis, where also variations in the mobile phase have to be considered. When large data sets with multiple compounds are compared with each other, matching of peaks between samples can be impossible without retention time alignment. Besides the linear shift correction *i*coshift (Tomasi et al., 2011) or the non-linear correlation optimized warping (COW) (Skov et al., 2006; Tomasi et al., 2004), many other algorithms for retention time alignment are available (Lange et al., 2007; Sinkov et al., 2011; Nielsen et al., 1998; Forshed et al., 2003; Szymańska et al., 2007; Walczak and Wu, 2005; Van Nederkassel et al., 2006). Many of the available software packages for non-targeted chromatographic data analysis align peaks after peak detection (for instance in XCMS).

For feature selection, peak picking and deconvolution of chromatogram segments of single samples are used. Most commonly, derivative based approaches are used (Felinger, 1998) to detect the location of peaks in a chromatogram, but a wide range of other methods are also available (Dixon et al., 2006; Furbo and Christensen, 2012; Hastings et al., 2002; Vivó-Truyols et al., 2005). Deconvolution techniques are only rarely implemented in chromatography software. Exceptions are for instance the freely available software Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein, 1999; Dromey et al., 1976) and the commercial software ChromaTOF (LECO, St. Joseph, MI, USA), which are often used in non-targeted chromatography studies (more on deconvolution in Chapter 2.2.4). For real chromatographic peaks and for deconvoluted peak profiles, either the peak height or the peak area is used as a quantitative measure. The final result for multivariate data analysis is a peak table. Some software packages report multiple entries per metabolite for peaks found for all m/z value, which can be problematic in multivariate analysis of the data (Behrends et al., 2011). Moreover, the quality of the final results are difficult

to evaluate because the identity of peaks are not known. For chromatographic preprocessing as well as for peak picking and deconvolution, visual examination of results is very important to avoid any introduction of artefacts by the used algorithms, albeit this validation can be time consuming and cumbersome. Moreover, a good system performance can often avoid the necessity to correct for noise, baseline deviations and peak shifts.

The final step of conventional non-targeted approaches is the explorative multivariate data analysis of the obtained peak table to reveal systematic structure and patterns in the data. Often univariate and multivariate statistical methods are used complementarily. T-tests, ANOVA or Fisher ratios (Pierce et al., 2006) are examples for univariate methods, which can be used to explore different levels of individual compounds across two or multiple groups of samples. The information from these univariate tests can for instance be used for variable selection prior to multivariate modelling. Multivariate approaches can be divided into supervised techniques, where classification groups are defined in advance, and unsupervised techniques, where classification groups are not known or can not be defined in advance.

Principal Component Analysis (PCA) is by far the most commonly used unsupervised method. PCA is a projection technique that searches for common patterns in a data matrix (e.g. peak table) to establish new directions explaining variance in the original data cloud. Onto these directions, called the loadings, each sample can be projected. These projections are called scores. A set of scores and loadings is a principle component (also called latent variable). The first principal component explains most of the variation in the data, while the explained variation decreases with the number of further principle components. PCA is often used to obtain a initial overview, as it can reveal unknown grouping of samples or confirm suspected groupings of samples. Another common unsupervised technique is Hierarchical Cluster Analysis (HCA). HCA first defines a clusters for each sample and successively

clusters samples together based on similarity measures until all samples constitute one cluster. The arrangement of the clusters (similarities among samples) are finally illustrated in a dendrogram (tree diagram) (Martens and Martens, 2001).

When group classifications are known in advance, more informative supervised multivariate models can be used. Supervised methods take, unlike PCA, intra-class variation (or within class variation) into account. PCA can however be coupled with the class information in order to give classification models by means of Soft Independent Modeling of Class Analogy (SIMCA), which is the first class modelling technique introduced in chemistry (Wold, 1976; Wold et al., 1981). SIMCA defines subspaces for each predefined class by providing a PCA for each class. A new sample is projected and compared to each subspace to evaluate its distance from the corresponding class. The assignment of the sample is done by comparing the distances of the sample from the class models. Another supervised method often used for classification of samples is Partial Least Squares Discriminant Analysis (PLS-DA). Partial Least Squares (PLS) was originally designed as a tool for statistical regression and became one of the most commonly used regression techniques in chemistry (Wold et al., 1966). In PLS-DA, the data matrix (peak table) is assigned as the independent variables (X-block) and the class coding is assigned as the dependent variables (Y-block). In a binary classification problem, classes in Y would be simply encoded as a class vector of ones and zeros. PLS-DA essentially searches for latent variables with a maximum covariance with the Y variables. Variation in the data matrix X which is not correlated with the class vector Y can affect the classification results. The interpretation of the results of PLS-DA can therefore significantly be improved by orthogonalizing the model (Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) (Trygg and Wold, 2002; Tapp and Kemsley, 2009)), which condenses the Y-block variance into the first latent variable.

2.2.3 Global multivariate modelling of chromatograms: Fingerprinting

Considering an elution profile as a sequence of single measured points, it appears obvious that chromatographic data have a very sparse nature. For instance, the quantitative information content of a chromatographic peak consisting of multiple data points can actually be expressed as a single number (e.g. peak area value or peak height). An ordinary chromatogram with a mass range between 50 and 300 m/z from a one-hour GC-MS run acquired at 4 Hz (four spectra per second) can be represented as a matrix of 14400 scans $\times$ 250 masses containing 3600000 data points. If we assume 200 peaks in this chromatogram, peak integration can reduce the data by 18000-fold! The actual content of information regarding chemical differences among samples in a set of chromatograms is even smaller. Moreover, information on systematic differences among samples in a peak table can often be decomposed into a few latent variables (principal components) using PCA.

Reduction and decomposition of information from chromatographic data can also be achieved by other means. Instead of using feature selection and multivariate modelling of a peak table, as in most software packages for non-targeted chromatographic analysis the chromatographic signal can be processed using mathematical techniques such as decomposition methods on a 'pixel-level', meaning chromatograms are processed in the format of raw detector data points.

The principle idea behind processing chromatograms as raw data points is the inclusion of as much information as possible to the multivariate analysis. In this way it can be avoided to set criteria where in the chromatogram chemically useful information is located (such as signal-to-noise ratio, peak width, peak shape and others), which is necessary when peak picking is applied. Consequently, important information can be missed using peak picking, as all peaks missing a certain criteria are simply not taken into account in further multivariate analysis. Automated peak integration can, depending on the degree of coelution and noise, also be troublesome

and often manual intervention is necessary.

Some strategies for modelling chromatographic raw data signals are reported. Most of them comprise common preprocessing tools such as baseline correction, noise reduction and peak alignment, as well as variable and/or data reduction techniques (Ballabio et al., 2008; Johnson and Synovec, 2002; Mohler et al., 2007; Borges, 2007; Sinkov et al., 2011; Sinkov and Harynuk, 2011, 2013; Teofilo et al., 2009; Pierce et al., 2005, 2006; Adutwum and Harynuk, 2014; Monforte et al., 2015; Jonsson et al., 2005; Bruce et al., 2008; Rodrigues et al., 2011; Silva Ferreira et al., 2014) or apply weights to variables (Christensen et al., 2005b,a; Christensen and Tomasi, 2007) prior to multivariate modelling. Similar concepts are also combined with multi-way analysis (Durante et al., 2011; Cocchi et al., 2008; Durante et al., 2006). Retention time alignment is the major disadvantage of all of these approaches, as alignment techniques are sometimes difficult to apply and prone to errors.

A small number of data processing approaches target a new representation of the chromatographic raw data signal by mathematical transformation to avoid retention time alignment of peaks among samples. These mathematical transformations include special correlation measures between data points (Danielsson et al., 2006), the calculation of R_V-coefficients (Daszykowski and Walczak, 2011), dissimilarity (Daszykowski et al., 2008) and distance matrices (Zerzucha et al., 2013) for two dimensional chromatograms of samples represented as matrices. The mathematical transformation used for these approaches eliminate information on the retention time of compounds and therefore hamper the identification of compounds responsible for differences between samples.

Another approach taking retention time shifts into account consists of segmentation of the two dimensional chromatograms along the retention time axis and deconvolution of the obtained chromatogram segments using a deconvolution method that takes retention times shifts into account. All samples of each segment are si-

multaneously deconvoluted using the two-way method Multivariate Curve Resolution (MCR) (Jellema et al., 2010) or the multi-way method Parallel Factor Analysis 2 (PARAFAC2) (Amigo et al., 2010a). Particularly, the simultaneous processing of samples ensures that no previous or subsequent alignment of retention times is necessary. Multivariate methods used for the deconvolution of coeluted chromatographic peaks are discussed more in detail in Section 2.2.4.

2.2.4 Local multivariate modelling of chromatograms: Resolution of peaks (Deconvolution)

Deconvolution is the mathematical resolution of overlapping peaks in a small section of the chromatogram (sometimes referred to as peak system). Peak profiles are estimated and pure spectra are obtained for identification. A perfect separation of peaks can not always be achieved, especially when very complex samples are analysed, or when fast chromatography is needed; in these cases deconvolution methods should be particularly favoured. Deconvolution techniques are mainly used in non-targeted data analysis approaches. They, however, can and should also be used to resolve overlapping peaks in conventional targeted analysis. As already mentioned in Chapter 2.2.2, few chromatographic software include deconvolution methods. Two examples are AMDIS (Stein, 1999; Dromey et al., 1976) and the commercial software ChromaTOF (LECO, St. Joseph, MI, USA). In contrast, high-level technical computing languages such as the freely available R or the commercial MATLAB offer versatile packages and toolboxes for multivariate modelling. Although R and MATLAB are command line driven programs, some packages provide graphical user interfaces (GUI). The great advantage of using computing languages for chromatographic data processing is the flexibility of combining functions from different packages and toolboxes.

Advanced factor models for the mathematical resolution of chromatographic peaks

(de Juan and Tauler, 2007; Amigo et al., 2010b; Brereton, 1995) and multivariate calibration models (Escandar et al., 2007; Ortiz and Sarabia, 2007) have been reviewed recently. When applied to chromatographic data, PCA is inadequate for finding direct chemically meaningful information, due to the rotational freedom of this bilinear model. This problem can be overcome with more advanced curve-resolution methods or factor models such as MCR-ALS (Tauler, 1995), Parallel Factor Analysis (PARAFAC) (Bro, 1997) and PARAFAC2 (Bro et al., 1999; Amigo et al., 2010a, 2008; Johnsen et al., 2014). Considering that each analyte has a distinct pattern, factor models are able to recover the elution and spectral profile. The following criteria must however be at least approximately true. Firstly, according to the Lambert-Beer law the collected spectra of a compound must behave linear to its concentration. Secondly, the intensity of each spectral point can be assumed to be the sum of the abundances of the analytes forming the mixture in each point of the elution profile. And lastly, the elution profile must be constant over samples. The shape and position of peaks between samples must not change. Note that PARAFAC2 and some special application of MCR take peak shifts and peak shape changes into account.

2.2.4.1 Multivariate curve resolution (MCR)

MCR is a bilinear model which is defined for a segment of a single chromatogram X, with J elution time points and K spectral points, as follows:

$$X = CS^T + E, \tag{2.1}$$

where X is a $J \times K$-matrix, C is a $J \times N$-matrix of elution profiles of N components, S^T is a $N \times K$-matrix of spectral profiles and E is $J \times K$-matrix of the residual error matrix.

Usually, an Alternating Least Squares (ALS) algorithm is used for MCR, which requires an initial guess of the number of eluting compounds (chemical rank of X).

This initial guess can be obtained by visual examination, singular values or PCA (de Juan and Tauler, 2007; Maeder and Zilian, 1988). Subsequently, an initial estimation of the spectral or concentration profiles for each compound obtained from e.g. Evolving Factor Analysis (EFA) is used to initialize the ALS procedure. Spectral and elution profiles are iteratively estimated under a series of constraints such as non-negativity, unimodality and sample selectivity (see de Juan and Tauler (2007); Bro (1998a) for detailed explanation of constrains) to decrease the extent of possible rotation ambiguities and give physical meaning to the obtained solutions. The algorithm stops when convergence criteria and constraints are met.

The above described deconvolution of a segment of a single chromatogram X can also be extended to multiple I samples X_i, as far as the number and the nature of the columns (spectra) are the same for all X_i matrices. The arrangement of the matrices and the extended bilinear model can be defined as:

$$X_{aug} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_I \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_I \end{bmatrix} S^T + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_I \end{bmatrix} = C_{aug} S^T + E_{aug}, \tag{2.2}$$

where X_{aug} is a $JI \times K$-matrix obtained by column-wise augmentation of all X_I data matrices. C_{aug} is a $JI \times N$ matrix of elution profiles of N components, S^T is a $N \times K$-matrix of the spectral profiles and the $JI \times K$ residual matrix E_{aug}.

2.2.4.2 Parallel factor analysis (PARAFAC)

PARAFAC is a multi-way decomposition method which, besides Tucker3, can be seen as a generalization of bilinear PCA to higher order data. PARAFAC can be expressed as a constrained version of Tucker3, and Tucker3 in turn as a constrained version of two-way PCA (Kiers, 1991). For the matrix x_{ij} and the three-way array

x_{ijk} the PCA model (Equation 2.3), TUCKER3 model (Equation 2.4) and PARAFAC model (Equation 2.5), respectively, are described as follows:

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij} \tag{2.3}$$

$$x_{ijk} = \sum_{f=1}^{F_1} \sum_{f=1}^{F_2} \sum_{f=1}^{F_3} a_{if} b_{jf} c_{kf} g_{f_1 f_2 f_3} + e_{ijk} \tag{2.4}$$

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{2.5}$$

Where F is the number of factors (components), a_{if}, b_{jf} and c_{kf} are elements of the loading matrices $A_{(I\times F)}$, $B_{(J\times F)}$ and $C_{(K\times F)}$. $g_{f_1 f_2 f_3}$ are the elements of the TUCKER3 core array, and e_{ij} and e_{ijk} are elements in the residual matrix $E_{(I\times J)}$ and residual array $\underline{E}_{(I\times J\times K)}$, respectively.

The PARAFAC model is visualized in Figure 2.2 and is written in matrix notation as

$$X_i = B D_i C^T + E_i \tag{2.6}$$

where X_i is the i-th frontal slab of the three-way array $\underline{X}$, D_i is a diagonal matrix with the i-th row of A in its diagonal and E_i residuals. C and B are loadings of the elution and spectral mode, respectively.

PARAFAC decomposes a three-way array into trilinear components. A component consist of three loading vectors, while no differentiation between scores and loadings

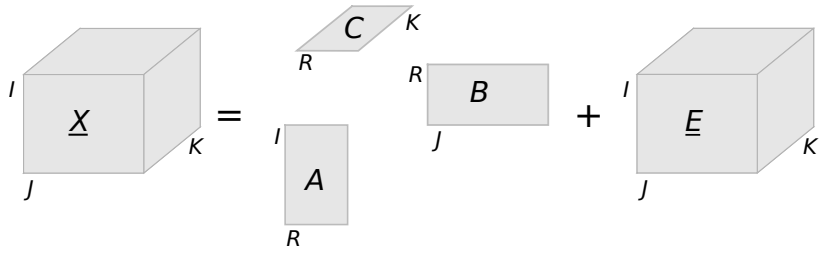


Figure 2.2: Visualisation of the PARAFAC model for a GC-MS data set $\underline{X}$ with I samples $\times$ J scans (elution profile) $\times$ K mass channels; the loading matrices A, B, and C; R factors (components) and the residual array $\underline{E}$.

is made in multi-way terminology. As PARAFAC requires low-rank trilinear data, chromatograms with many peaks can not be resolved in entirety and must be split into segments containing only a few peaks (Bro et al., 2001; Amigo et al., 2008). Unlike in a bilinear PCA, extracted PARAFAC components are not orthogonal. Consequently, extracted components are allowed to relate to each other as long as the difference in the elution and spectral profile is big enough and can be identified as individual contributions to the overall signal. Moreover, the PARAFAC model is not nested, which means that for instance the first component of a two component model does not reflect the same information as a one component model. As a PARAFAC model can not be rotated without a loss of fit (no rotational freedom), only one unique, best-fit solution is possible with a certain number of components. It is therefore essential to determine the proper number of components for a PARAFAC model. Due to this uniqueness, chemically meaningful elution and spectral profiles are provided when a small region of a chromatogram is deconvoluted.

2.2.4.3 Parallel factor analysis 2 (PARAFAC2)

PARAFAC2 is a constrained version of PARAFAC, which can handle unsystematic retention time shifts of peaks between samples and to a certain degree changes in peak shapes. Unique and chemically meaningful solutions can be obtained while the

natural data structure (shifting peaks) of a set of two-dimensional chromatograms is taken into account. The elution profile is accepted to change among samples (e.g. peaks shift, distortion of peak shape, different length) as long as the cross product (sums of squares and cross products) of the elution profile remains constant over all samples. The PARAFAC2 model is visualised in Figure 2.3 and can be written in matrix notation for the decomposition of an $I \times J \times K$ three-way array $\underline{X}$ as

$$X_i = B_i D_i C^T + E_i = (P_i) D_i C^T + E_i \qquad \forall i = 1, \ldots, I, \tag{2.7}$$

where X_i is the i-th frontal slab of the three-way array $\underline{X}$, D_i is a diagonal matrix with the i-th row of A in its diagonal and E_i the residuals. C is the loading matrix for the spectral mode and B_i the loading matrix of the elution mode for the i-th slab of $\underline{X}$ modeled as $P_i H$. For F factors (or components) P_i is an $I \times F$-matrix and H is a $F \times F$-matrix. P_i and H have no direct chemical interpretation, but their product is an estimate of the elution profiles B_i.

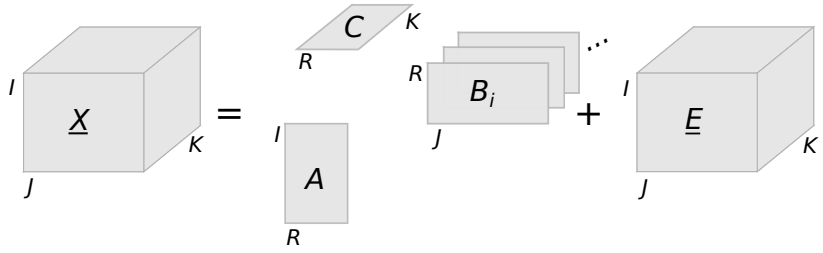


Figure 2.3: Visualisation of the PARAFAC2 model for a GC-MS data set $\underline{X}$ with I samples $\times$ J scans (elution profile) $\times$ K mass channels; the loading matrices A, B_i, and C; R factors (components) and the residual array $\underline{E}$.

2.3 Wine aroma

Wine aroma is the perceived odour of a mixture of volatile wine constituents through the human nose. The sensory characteristics which, in fact, are the major quality criterion of a wine, essentially depend on the volatile composition. Dedicated information on volatile constituents in wine is important to the winemaker aiming to produce a product fulfilling consumer sensory expectations. The modulation of wine aroma presupposes a broad understanding of the impact of the different steps of vinification on the composition of aroma compounds (Bisson et al., 2002; Swiegers et al., 2005).

The word aroma refers to the smell of a wine. Wine aroma can be distinguished into primary, secondary and tertiary aromas. The primary aromas contributing to the varietal character originate from the grapes. The secondary aromas derive from alcoholic and malolactic fermentation and the tertiary aromas are formed during the maturation and ageing process in the barrel and bottle. The term *bouquet* refers to aromas evolving during maturation in the bottle. The term flavour includes the aroma and taste (sweetness, bitterness, acidity, saltiness, umami) of a wine and is often incorrectly interchanged with the term aroma in popular usage (Swiegers et al., 2005; Clarke and Bakker, 2004). Natural products such as wine often contain hundreds of volatile compounds with different properties regarding their odour potentials. The sensory threshold is a very important characteristic of a volatile compound. In complex mixtures the odours of compounds may stay distinct, suppress each other or synergistically create another sensory impression. Even non-volatile compounds or compounds present below their threshold levels can therefore affect the perceived aroma of wine.

The volatile composition of wine consists of several hundreds of compounds. Not all, but a large number of these compounds which originate either from the grapes, wine microbes or the maturation process (wood derived substances from barrels)

contribute to wine aroma. The alcoholic fermentation with yeast is particularly important in the formation of wine aroma. The production of major and minor odour active metabolites by yeast from for instance sugar and amino acids, and the conversion of grape derived non-volatile precursor such as glyco- and cystein conjugated compounds are crucial for the development of general wine aroma and in some wines for the varietal aroma character. Lactic acid bacteria used for Malolactic Fermentation (MLF) after or during alcoholic fermentation also influence wine aroma, albeit to a lesser extent than yeast. Lactic acid bacteria also produce aroma active metabolites through the conversion of compounds derived from grapes or alcoholic fermentation. The major goal of MLF is the reduction of acidity by the conversion of harsh tasting L-malic acid to milder tasting L-lactic acid. Consequently, the style of a wine can be significantly influenced by MLF (Swiegers et al., 2005; Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000).

2.3.1 Volatile wine compounds

The most important groups of volatile compounds found in wine are discussed in the following sections.

2.3.1.1 Alcohols

Varying concentrations of ethanol in wine between 7 and 16 % (v/v) have an impact on the solubility and volatility of aroma active compounds (Yu and Pickering, 2008). Consequently, the ethanol content influences the sensory perception of a wine. Ethanol plays an important role in the formation of ethyl esters. Methanol, well known for its toxicity, occurs only in very low quantities in wine and originates solely from enzymatic degradation of grape pectin (Ribéreau-Gayon et al., 2000).

Higher alcohols, sometimes also referred to as fusel alcohols, are aliphatic or branched alcohols with more than two carbons and make up for the largest quantity

of volatiles in wine (Ribéreau-Gayon et al., 2000; Sumby et al., 2010). These alcohols with higher molecular weights and higher boiling points are produced by yeasts from amino acids via the Ehrlich pathway or from sugars. For example, a higher alcohol can be related to an corresponding amino acid such as 3-methylbutanol from leucine, 2-methylbutanol from isoleucine and 2-methylpropanol from valine. Concentrations produced during fermentation depend on many different factors such as the yeast species and strain, composition of nitrogen containing compounds of the must (e.g. amino acids, ammonia), pH, oxygen levels and temperature. The concentration of these alcohols determine their impact on the sensory perception of wines. Low concentrations ($< 300\,\mathrm{mg\,L^{-1}}$) can contribute to the complexity of a wine, while higher levels lead to pungent odours, suppressing the fruitiness and elegance of a wine (Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000).

Another important group of alcohols are the C6-alcohols such as hexanol and cis-3-hexenol. These compounds are associated with green, herbaceous notes. C6-alcohols occur in high concentrations in wines made from unripe grapes. The corresponding aldehydes of these alcohols are degradation products of linoleic and linolenic acids (Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000). The unsaturated secondary alcohol 1-octen-3-ol is particularly found in botrytized wines (Ribéreau-Gayon et al., 2000; Rapp and Mandery, 1986). This compound has an odour of mushrooms and is a well-known fungal metabolite also related to molds such as *Aspergillus* and *Penicillium* (Kaminski et al., 1974).

2.3.1.2 Esters

Esters are the primary source of fruity aromas and the second major constituents of wine volatiles after fusel alcohols. The composition of esters and synergistic effects influence various fruity notes. Esters are therefore also very important for the overall sensory perception of a wine (Lytra et al., 2012, 2013). Wine esters are either formed

enzymatically or evolve during wine ageing by chemical esterification of alcohols and acids. Enzymatic ester synthesis by yeast is catalysed by esterases, lipases and alcohol acetyltransferases (Sumby et al., 2010).

The quantity of esters formed during fermentation depends on the activity of the involved enzymes, the yeast strain, nutrition status, fermentation temperature, and the degree of must clarification. Esters which were produced during fermentation in excess of their equilibrium hydrolyse during wine ageing, as the chemical esterification and hydrolysis of ester is an equilibrium reaction. Ester hydrolysis is favoured at high temperature and low pH. In fact, depending on this reaction the equilibrium levels of some esters increase during wine ageing. Branched fatty acid ethyl esters tend to increase as a function of time, since they are present at low levels after fermentation (Ribéreau-Gayon et al., 2000; Sumby et al., 2010).

The possible variety of esters is enormous considering the large number of different acids and alcohols in wine. The wide range of esters in wine can be grouped according to similar structure or physiochemical properties as follows: major aliphatic ethyl esters (even number of carbons), aliphatic ethyl esters (odd number of carbons), ethyl esters of branched aliphatic acids, aromatic esters, acetates of higher alcohols, methyl esters, minor isoamyl esters, and others (Antalick et al., 2010b).

2.3.1.3 Fatty acids

Acids contributing primarily to the titratable acid of wine namely tartaric acidity, malic acid and lactic acid are not volatile. The concentrations of these acids can however impact the aroma by playing a role in the release of aroma compounds from wine. Volatile acidity (VA) consists of approximately 90 % acetic acid. Yeast produces olfactorily imperceptible amounts of acetic acid. Perceptiple amounts of acetic acid can however originate from microbial spoilage, in particular from some lactic acid bacteria and *acetobacter* species. Moreover, increased levels of propanoic

acids, butanoic acids and especially 3-methylbutanoic acid (isovaleric acid) are associated with microbial contamination. Hexanoic, octanoic and decanoic acid derive from yeast metabolism. In high concentrations, these compounds can lead to rancid, pungent, cheese and fat-like odours and are considered to cause stuck fermentations (Swiegers et al., 2005; Ribéreau-Gayon et al., 2000; Francis and Newton, 2005).

2.3.1.4 Carbonyl compounds

Aldehydes are oxidation products of primary alcohols. Acetaldehyde (ethanal) is the most abundant carbonyl compound in wine. The formation of acetaldehyde occurs during alcoholic fermentation and depends mainly on must composition, must clarification and aeration status. Moreover, acetaldehyde can increase over time due to oxidation of ethanol and activity of spoilage yeast (Bennetzen and Hall, 1982; Denis et al., 1983; Fleet, 1993). Aldehydes in general react with sulphur dioxide (formation of bisulfite adducts). Consequently, insufficient addition of sulphur dioxide during the wine making process leads to elevated levels of free acetaldehyde, which is negatively perceived as 'flatness'. Aldehydes are also associated with oxidized aroma notes in wines, such as 'cut-apple' and 'nutty' odours. During vinification acetaldehyde also plays an important role as an binding partner for phenolic compounds and has therefore an impact on the formation of color pigments and tannins (Boulton, 2001; Timberlake and Bridle, 1976). Analogous to the C6-alcohols, C6-aldehydes such as hexanal and cis-3-hexenal contribute to 'green', 'herbaceous' odours. Aromatic, wood derived aldehydes such as vanillin and cinnamic aldehyde can contribute to tertiary aromas of wine.

Ketones are oxidation products of secondary alcohols. The most important compound in this class formed in wine is the diketone diacetyl (2,3-butandione). While yeast is responsible for the production of large amounts of diacetyl during beer fermentation, lactic acid bacteria are the main source of this vicinal diketone in wine,

albeit wine yeasts also produce insignificant amounts of this compound. Malolactic fermentation can be conducted in a controlled manner, but undesired activity of spontaneous lactic acid bacteria flora can lead to spoilage of the wine. The sensory impact of diacetyl in wine is described as sweet, buttery and butterscotch. These odours are perceived as pleasant in low concentrations, higher concentrations however, lead to an objectionable off-flavour. Diacetyl production of lactic acid bacteria during malolactic fermentation can be controlled by several factors such as the malolactic bacteria strain, inoculation dosage, temperature, pH, citric acid content and sulphur dioxide concentrations used during vinification. The latter results from the above mentioned reaction of carbonyl compounds with bisulfid ions. Diacetyl can be reduced to 2,3-butanediol in wine conditions. 2,3-butanediol has a much higher odour threshold then diacetyl, which is rarely exceeded in wine (Bartowsky and Henschke, 2004; Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000).

2.3.1.5 Terpenes

C5-Isoprene units are the building blocks of terpenes. The most important classes of terpenes are the monoterpenes consisting of two isoprene units, sequiterpenes consisting of three isoprene units and the C13-norisoprenoids. Chemically modified terpenes through oxidation or rearrangement are called terpenoids. In the following discussion the term terpene will be used to include all terpenoids for the sake of simplicity (Ribéreau-Gayon et al., 2000).

A large number of monoterpenes and monoterpene derivatives containing alcohol (e.g. linalool), aldehyde (e.g. geranial), acid (e.g. *trans*-geranic acid) and ester groups (e.g. geranyl and neryl acetate) have been reported in wine. Linalool, α-terpineol, nerol, geraniol, citronellol and hotrienol are the most important compounds of this group due their relatively low olfactory thresholds (in the $\mu g\,L^{-1}$ range). Terpenes, which mainly derive from grapes, are responsible for the aroma of Muscat wines such

as Muscat d'Alsace, Muscat á Petits Grain and Muscat d'Alexandria. Terpenes are also responsible for the 'Muscat-like' characteristics of aroma related cultivars such as Gewürztraminer, Riesling and Scheurebe, commonly grown in Germany and Alsace, France. Terpenes may contribute to the aromas of non-muscat varieties as well. In other very popular grape cultivars such as Sauvignon blanc, or particularly red varieties such as Cabernet Sauvignon, Merlot, Cabernet franc and Syrah, terpenes are usually present under their olfactory thresholds and do therefore not play a significant role in the aromas of these cultivars (Marais, 1983; Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000).

A large extent of terpenols (including diols and triols) in grapes are bound as non-volatile glycosides and are therefore not aroma active. These glycosides mainly contain the monosaccharide β-D-glucose and the disaccharides α-L-arabinofuranose-β-D-glucopyranose, α-L-rhamnopyranose-β-D-glucopyranose, β-D-xylopyranose-β-D-glucopyranose and β-D-apiofuranose-β-D-glucopyranose. Besides terpenols, other compounds with hydroxyl groups such as hexanol, 2-phenyl ethanol, benzyl alcohol, C13-norisoprenoids and volatile phenols (e.g. vanillin) are present in glycosylated from. Due to higher water solubility, glycosides serve as carriers for the transport and accumulation of the corresponding aglycones in plants. Muscat grape varieties have a particularly large ratio of glycosylated terpenols to free from, whereas this ratio for non-muscat cultivars is approximately 1:1. Aglycones of glycosides can be realeased either enzymatically or by acid hydrolysis, whereas the latter plays a minor role in wine. Enzymes with glycosidic activity responsible for the liberation of aroma compounds are mainly sourced from yeasts, but also bacteria and grapes. Oenological enzymes used for clarification can also have glycosidic side activity (Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000; Black et al., 2015).

C13-norisoprenoids are degradation products of carotenoids and can be grouped into megastigmanes and non-megastigmanes. Two examples of megastigmanes with

a very low perception threshold of only several $\mu g\,L^{-1}$ are β-damascenone and β-ionone, which contribute to 'fruity' and 'flowery' notes in wine (Sefton et al., 1989; Mendes-Pinto, 2009). The most important non-megastigmane is 1,1,6-trimethyl-1,2-dihydronaphtalene (TDN) which is responsible for the distinct 'kerosene' odour in Riesling and contributes to the ageing bouquet of Riesling wines (Winterhalter et al., 1990; Winterhalter, 1991).

2.3.1.6 Sulphur and nitrogen containing compounds

The majority of volatile sulphur containing compounds in wine are associated with reductive off-flavours. Some thiols however have positive sensory characteristics and contribute to the varietal aroma of certain grape varieties. Negatively perceived volatile sulphur compounds are either directly or indirectly linked to yeast metabolism. Residues from sulphur containing spray agents and thermal or photochemical reactions can also be a source of volatile sulphur compounds. Volatile sulphur compounds are often divided into low-boiling and high-boiling compounds (Swiegers et al., 2005).

High concentrations of the low-boiling sulphur compounds methanethiol, ethanethiol and particularly hydrogen sulphide lead to reductive off-flavours such as 'rotten egg' and 'sewage'. Hydrogen sulphide is a yeast metabolite formed intracellularly by the reduction of sulphates and the metabolisation of sulphur containing amino acids such as cysteine and methionine. The production of hydrogen sulphide is therefore strongly linked to nitrogen metabolism. High production of hydrogen sulphide due to nitrogen deprivation during fermentation can be avoided by the addition of ammonium sulphate in the early stages of fermentation. Hydrogen sulphite can react with methanol and ethanol to produce methanethiol and ethanethiol (Lambrechts and Pretorius, 2000; Swiegers et al., 2005). Dimethyl sulphide (DMS) is a rare example of a positively associated low-boiling volatile sulphur compound. This sulphur compound

is formed by yeast, it evolves during ageing and is therefore considered to contribute to the bouquet (De Mora et al., 1986; Silva Ferreira et al., 2003; Picard et al., 2015).

High boiling sulphur volatiles are only of minor importance to wine aroma, although methionol is an exception. The deamination and decarboxylation of methionine according to the Ehrlich pathway, results in the formation of methionol, which is perceived as 'cauliflower' aroma in higher concentrations (Ribéreau-Gayon et al., 2000).

The varietal thiols 4-mercapto-4-methyl-pentan-2-one (4MMP), 3-mercaptohexan-1-ol (3MH), 3-mercapto-3-methyl-butan-1-ol (3MMB), 4-mercapto-4-methyl-pentan-1-ol (4MMPOH) and 3-mercaptohexanolacetate (3MHA) have been identified as key molecules in some grape varieties. The varietal aroma of Sauvignon blanc is particularly determined by these thiols (besides the methoxypyrazines). Other varieties such as the white cultivars Semillon, Scheurebe and Riesling, or the red cultivars Cabernet Sauvignon, Merlot and Pinot noir among others also contain varying amounts of these thiols. The single compounds have different odour expressions of 'boxtree' and 'passion fruit' (4MMP); 'passion fruit', 'grapefruit', 'gooseberry' and 'guava' (3MHA & 3MH); and 'cooked leeks' (3MMB) (Swiegers et al., 2009; Roland et al., 2011). Similar to terpenols, these thiols result from the cleavage of odourless precursors by yeast enzymes during alcoholic fermentation, whereas the nonvolatile precursors are not glycosides, but S-cysteine conjugates. It is assumed that yeast originated β-lyases are responsible for the non-quantitative release of these thiols during fermentation (Tominaga et al., 1998; Peyrot des Gachons et al., 2000). Alternatives to the classical pathway from cysteine conjugates, such as the 1,4-addition of hydrogen sulphide to conjugated carbonyl compounds (e.g. *E*-hex-2-enal), have also been described (Schneider et al., 2006).

With few exceptions, volatile nitrogen compounds are of minor importance regarding the aroma of wine. In Cabernet Sauvignon, Sauvignon blanc and Caber-

net franc, the grapevine metabolites 3-alkyl-2-methoxypyrazines matter particularly. The compounds 3-isopropyl-2-methoxypyrazine, 3-isobutyl-2-methoxypyrazine and 3-*sec*-butyl-2-methoxypyrazine have very low perception thresholds and are the most studied in this group contributing to aromas of 'green bell pepper', 'asparagus' and 'earthy'. Undesired herbaceous notes in Cabernet Sauvignon and Cabernet franc wines made from unripe grapes are attributed to 2-methoxy-3-isobutylpyrazine. 2-methoxy-3-isobutylpyrazine is located in the grape skins and therefore increases during fermentation and maceration. On the other hand, herbaceous notes associated with 2-methoxy-3-isobutylpyrazine such as 'green bell pepper' can be desirable in Sauvignon blanc wines (Allen et al., 1991; Lacey et al., 1991; Ribéreau-Gayon et al., 2000).

Some thiazoles and oxazoles are thought to contribute to the ageing aroma of wine. Athough the mechanisms of the formation of these compounds are not yet fully understood, some might be formed in a Maillard-type reaction between carbonyl or dicarbonyl compounds and amino acids. (Keim et al., 2002; Marchand et al., 2000, 2002, 2011)

2.3.1.7 Other volatile compounds

Lactones and furans are compounds of different origin which influence wine aroma. Lactones are formed by intra molecular condensation of a hydroxy and a carboxy group resulting in an cyclic ester. Saturate γ-lactones are also called dihydrofurans. Lactones can be arise during fermentation by rearrangement of hydroxycarboxylic acid obtained from deamination and decarboxylation of amino acids. Some lactones are associated to specific grape varieties. For instance, 2-vinyl-dihydrofuran-2-one is present in Riesling and Muscat wines and 2,5-dimethyl-4-hydroxy-3(2H)-furanone (furaneol) can be found in Merlot and Vitis lambrusco wines. The sotolon (3-hydroxy-4,5-dimethyl-2(5H)-furanone) is linked to botrytized and fortified wines and marker

for premature oxidative ageing of wine. Sotolon can be formed by condensation of α-keto butyric acid and acetaldehyde. The 'oak lactones' or 'whiskey lactones', which are the *cis*- and *trans*-isomers of 3-methyl-γ-octalactone, contribute to the 'oaky' aroma of wines vinified in barrels. Other compounds of this class may arise from saccharide degradation and through the Maillard reaction (Muller et al., 1973; Clarke and Bakker, 2004; Ribéreau-Gayon et al., 2000).

Another group of important wine compounds are volatile phenols. The four compounds 4-vinylphenol, 4-vinylguaiacol, 4-ethylphenol and 4-ethylguaiacol are predominantly associated with objectionable 'phenolic' character. 4-ethylphenol and 4-vinylphenol are related with odour descriptors as 'barnyard', 'sweaty saddle' and 'medicinal', 'Band Aid', respectively. These odours are mainly perceived as unpleasant, while 4-vinylguaiacol and 4-ethylguaiacol have positive odours of carnations and 'smoky', 'spicy', respectively (Chatonnet et al., 1997). These compounds are formed through enzymatic degradation of the cinnamic acids *p*-coumaric and ferulic acid by yeast (*Saccharomyces cerevisiae*) derived cinnamate decarboxylase. Other phenolic compounds such as procyanidins inhibit cinnamate decarboxylase activity resulting in lower levels of 4-vinylphenols in red wines compared to white wines. The concentration of this compound in white wine depends on the activity of cinnamate decarboxylase and concentration of the precursors, which in turn vary among grape cultivars (Chatonnet et al., 1997; Du Toit and Pretorius, 2000; Ribéreau-Gayon et al., 2000).

Volatile phenols can also derive from spoilage by *Brettanomyces/Dekkera* yeasts, which express a cinnamate decarboxylase which is not inhibited by phenolic compounds resulting in the conversion of large quantities of cinnamic acids to 4-vinylphenol and 4-vinylguaiacol. These spoilage yeasts also produce vinylphenol reductase, which is absent in *Saccharomyces cerevisiae*, catalysing further reduction of 4-vinylphenol and 4-vinylguaiacol to 4-ethylphenol and 4-ethylguaiacol. Proper sulphur dioxide

management during vinification can prevent growth of these spoilage yeasts (Chatonnet et al., 1995, 1997; Ribéreau-Gayon et al., 2000).

2.3.2 Gas chromatography in wine analysis

Wine aroma is the perceived scent of wine, which in turn is the detection of volatile wine constituents by means of the olfactory nerves in the human nose. Gas chromatography is the most suitable analytical technique for the analysis of volatile compounds and therefore, the most widely used method for the analysis of aroma compounds in wine. Targeted analysis of wine aroma compounds are commonly conducted. A targeted approach always presuppose an *a priori* defined set of compounds of interest. In numerous domains of analytical sciences including wine analysis, non-targeted strategies have recently gained more attention. Non-targeted approaches focus on the extraction and analysis of as many compounds as possible in the analysed samples to obtain a more comprehensive picture of the sample composition. Targeted and non-targeted approaches to gas chromatography have been more generally discussed in Sections 2.1 and 2.2, respectively. In the next section, targeted and non-targeted analysis is specifically discussed in a wine context.

2.3.2.1 Conventional targeted analysis of wine volatiles

In principle, all commercially available separation columns, injection systems, and detectors are used for the analysis of wine volatiles using gas chromatography. Commonly more polar column phases such as polyethylene glycol (PEG) or modified PEG (WAX) (Ferreira et al., 1993; Bonino et al., 2003; Boido et al., 2009; Ugliano and Moio, 2005) are preferred due to the diverse nature of volatile wine constitutes, but also non-polar phases such as polydimethylsiloxane (PDMS) (Escudero et al., 2007; Sánchez-Palomo et al., 2005) or enantio-selective (cyclodextrin based) phases for chiral separations are used (Fernandes et al., 2003).

Sample preparation is in general a crucial point in GC analysis. For the analysis of wine volatiles interfering matrix constituents such as water, alcohol and non-volatiles have to be taken into account. The selection of a sample preparation technique depends mainly on the physiochemical properties (e.g. polarity) and the concentration of analytes. Aroma compounds in wine are often loosely differentiated between major and minor volatiles. Major volatiles are mainly higher alcohols, some esters and fatty acids. Liquid-Liquid Extraction (LLE) with for instance diethyl ether (Louw et al., 2006; Lilly et al., 2000), dicloromethane (Selli et al., 2006; Perestrelo et al., 2006; Mallouchos et al., 2003) or Freon 113 (Ferreira et al., 1993; Muñoz et al., 2007) and Solid Phase Microextraction (SPME) are the most commonly used sample preparation techniques for these compounds present in high concentrations in wine. The analysis of minor volatiles can be very difficult in terms of the extraction, enrichment and detection of analytes. Solid Phase Extraction (SPE) meets these requirements for the trace analysis of minor compounds, as it is applicable to a wide range of compounds due to the availability of different commercial phases. Usually, reversed-phase C18 (Lukić et al., 2006), Lichrolute EN (Loscos et al., 2007; Lopez et al., 2002) and styrene diviniylbenzene phases (Palomo et al., 2005) are used.

The major problem with LLE and SPE are hazardous properties of organic solvents used which determined by their molecular structure can be toxic, flammable, carcinogenic and/or neurotoxic. Furthermore, all organic solvents are environmentally hazardous, especially the greenhouse gas, Freon. Solvent free techniques are therefore preferred and gain more and more popularity. SPME is a solvent free and fully automatable alternative to LLE and SPE. Fibres with different characteristics have been used for the analysis of wine volatiles namely: PDMS (Riu-Aumatell et al., 2006; Alves et al., 2005), carboxen/PDMS (CAR/PDMS) (Piñeiro et al., 2006), PDMS/Divinylbenzene (PDMS/DVB) (Sánchez-Palomo et al., 2005), DVB/CAR/PDMS (Sánchez-Palomo et al., 2005), polyethyleneglycol/DVB (PEG/DVB) (Flamini et al.,

2005) and polyacrylate (De la Calle García et al., 1997). A more recent development of a solvent free sample preparation technique suitable for the analysis of wine volatiles is stir bar sorptive extraction (SBSE) (Hayasaka et al., 2003; Zalacain et al., 2007; Weldegergis and Crouch, 2008). SBSE shows significant increase in sensitivity compared to SPME due to the higher phase volume, and is therfore also suitable for the analysis of trace compounds (Sandra et al., 2001; Zalacain et al., 2004). Besides PDMS, a more polar mixed phase of ethylene glycol and PDMS is available since recently, which facilitates the extraction of more polar wine volatiles (Elpa et al., 2014).

2.3.2.2 Non-targeted approaches to wine volatiles

Inspired by the new field of metabolomics the number of wine related studies comprising non-targeted strategies have steadily increased in the last years. A variety of methodologies for non-targeted analysis of wine volatiles using GC-MS have been applied to several oenological and viticultural questions. A review outlining a variety of reported studies on wine metabolite profiling is given by Atanassov et al. (2009). To give a brief overview on the applicability of non-targeted GC-MS analysis to wine, some of the more recent publications are summarised in the following.

Castro et al. (2012) used among other techniques non-targeted HS-SPME-GC-MS analysis in combination with the software package MetAlign (Lommen and Kools, 2012) to study the effect of oxidative response of *Saccharomyces cerevisiae* during fermentation. Castro et al. (2014) developed a process analytical technology pipeline including the combination of GC-MS data preprocessing and multivariate analysis to investigate 'forced ageing' of Port wine. Another non-targeted study on volatiles related to port wine aging from Jacobson et al. (2013) uses GC-FID, multivariate statistics and network reconstruction. Network reconstruction of preprocessed GC-MS data has also been used by Monforte et al. (2015) to study kinetics of port

wine aging. A methodology by Schmidtke et al. (2013) uses multivariate curve resolution applied to GC-MS profiles coupled with full descriptive sensory analysis to determine the objective composition of various styles of Australian Semillon wines. Robinson et al. (2011a,b) developed a non-targeted method for characterizing the wine volatile profile using Headspace Solid Phase Microextraction Comprehensive Two-dimensional Gas Chromatography Time-of-flight Mass Spectrometry (HS-SPME-GC×GC-TOFMS) and studied the influence of yeast strain, canopy management, and site on the volatile composition and sensory attributes of Cabernet Sauvignon wines. A non-targeted strategy for the varietal authentication of German white wines based on Headspace Solid Phase Microextraction Gas Chromatography Mass Spectrometry (HS-SPME-GC-MS) and multivariate classification was published by Springer et al. (2014). Fedrizzi et al. (2012) introduced an optimization procedure for non-targeted HS-SPME-GC-TOF metabolite profiling of grape volatiles using D-optimal design.Howell et al. (2006) used a non-targeted GC-MS method to show that multiple strains of *Saccharomyces* grown together in grape juice can affect the profile of aroma compounds that accumulate during fermentation. Silva Ferreira et al. (2014) describe a non-invasive, high throughput GC-MS methodology facilitating 'real time' monitoring of the metabolic changes during fermentation of *Saccharomyces cerevisiae* in synthetic grape must containing different sources of yeast assimilable nitrogen. A study conducted by Conterno et al. (2013) used non-targeted and targeted metabolomic approaches to reveal compounds which characterise the growth of *Dekkera bruxellensis* in media with low nutrient availability and different ethanol concentrations modelling the wine environment. In a study on lactic acid bacteria, Lee et al. (2009) compared the metabolic profile of isolated *Lactobacillus plantarum* and commercial *Oenococcus oeni* using GC-MS and Nuclear Magnetic Resonance (NMR). The combination of non-targeted GC-MS and NMR analysis was also used to unravel metabolites in grape juice that affect the production of varietal thiols in Sauvignon

blanc wines by Pinu et al. (2014).

Besides GC-MS other analytical techniques have been used for non-targeted analysis of wine constituents. As in classical metabolomics, LC-MS (Tarr et al., 2013; Arapitsas et al., 2016, 2014, 2012; Roullier-Gall et al., 2015; Arbulu et al., 2015; Toffali et al., 2011) and NMR (Lopez-Rituerto et al., 2012; Laghi et al., 2014; Rochfort et al., 2010) are commonly used analytical technique. These techniques are here, however, not further discussed.

2.4 Rapid sensory profiling of wine

The objective of sensory profiling is to provide a visualisation of differences between samples perceived by a taster in the form of a product map. The quality of these maps depend on certain criteria such as the repeatability of blind duplicates, representation of descriptive attributes of the samples, interpretability and clear representation of the results. The outcome should be useful to either confirm a hypothesis or postulate a new hypothesis. Sensory profiles of multiple samples meeting these aforementioned requirements can be obtained using conventional sensory profiling methods, such as Quantitative Descriptive Analysis (QDA) (Stone et al., 2008), in combination with multivariate data analysis (e.g. PCA). These conventional techniques, however, require intensive training of panellists and are therefore time-consuming.

Rapid descriptive methods provide a view on the sensory differences among samples similarly to conventional profiling methods. Labour-intensive panel training is, however, omitted or reduced, resulting in a dramatic decrease of the total analysis time (Risvik et al., 1994, 1997). This saving of time comes with the cost of sacrificing quantitative data of defined sensory attributes. When defined sensory descriptors as in conventional sensory profiling are used, information on the importance of other/different attributes in the overall perception of panellists is not obtained. Some rapid sensory profiling methods overcome this problem by allowing the taster to more freely decide how to indicate differences between samples. These faster alternatives, such as perceptual mapping (e.g. 'napping') with Ultra Flash Profiling, provide citation frequencies of sensory descriptors freely chosen by the assessors, which explain sensory differences in the sample set (Pagès, 2005a; Delarue and Sieffermann, 2004; Cartier et al., 2006; Dehlholm et al., 2012). The increasing number of publications on the application of rapid sensory profiling techniques to food stuffs and beverages testify that these methods have recently gained more popularity (Perrin et al., 2008;

Nestrud and Lawless, 2008, 2010; Kennedy, 2010; Ross et al., 2012; Torri et al., 2013; Santos et al., 2013).

Two recent reviews give a comprehensive overview on theoretical background, implementations, advantages and disadvantages and comparison of different rapid descriptive methods (Varela and Ares, 2012; Valentin et al., 2012). The expeditious means of these novel sensory profiling techniques are very well elucidated by the expressive title of one of these reviews: *Quick and dirty but still pretty good: a review of new descriptive methods in food science* (Valentin et al., 2012). In the following a brief overview on rapid descriptive methods is provided, which can basically be defined into three groups. The first group are verbal-based methods including free choice profiling (Williams and Langron, 1984), flash profiling (Dairou and Sieffermann, 2002) and check-all-that-apply questionnaires. From verbal-based methods a direct description of the products is obtained similar to QDA, but the time-consuming steps of attribute and scaling alignment of classical methods is avoided. The second group are reference-based methods, which include preselected reference sample. Polarised sensory positioning (Teillet et al., 2010) and pivot profiling (Thuillier et al., 2015) are examples for this group. The third group consists of similarity based methods which focus on the overall assessment of the similarity of samples. The most important techniques belonging to the third group are sorting (Lawless et al., 1995; Schiffman et al., 1981), projective mapping (Risvik et al., 1994) and its modification napping (Pagès, 2003) on, which the main emphasis is laid in the following.

2.4.1 Projective mapping

During projective mapping, assessors are encouraged to position a set of samples on a sheet of paper according to perceived similarities. Samples which are perceived as similar are placed close to one another and samples which are perceived as different, are positioned away from one another. An fictitious example of a taster sheet is shown

in Figure 2.4. X- and y-coordinates for each sample are collected and summarized in a table for each taster. Panellists can also be asked to describe each product by writing a few words (freely chosen) directly on the sheet near the products, which has been referred to as Ultra Flash Profiling (Perrin et al., 2008). Assessors are permitted to re-taste the samples as often as they want and to take as much time as needed. The sensory attributes provided for each wine are collected, similar descriptors are usually grouped together, and the citation of each descriptor group is finally counted for each wine. In this manner a table of citation frequencies of each descriptor group for each sample is obtained. Usually not more than 10 to 15 samples can be evaluated depending on the product and how pronounced the differences among samples are. The final structure of the data is displayed in Figure 2.5. The napping approach is a special way of performing projective mapping with a specified protocol regarding paper size, task instructions and data analysis method (MFA). Different modification of the napping approach have been reported (Pagès, 2005b; Perrin et al., 2008; Perrin and Pagès, 2009; Pagès et al., 2010).

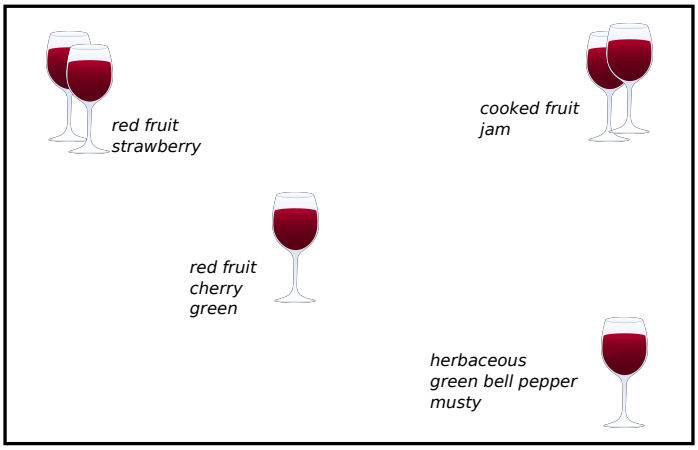


Figure 2.4: Fictitious example of a projective mapping sheet of six red wines with freely chosen sensory descriptors from Ultra Flash Profiling.

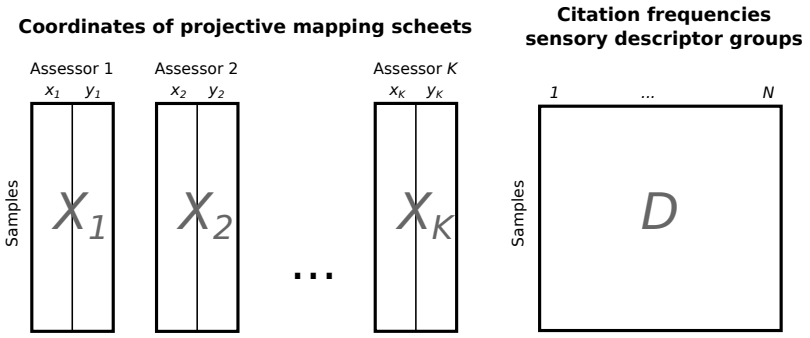


Figure 2.5: Data structure of projective mapping with Ultra Flash Profiling. Tasting sheets of K assessors are represented as matrices X_k which consist of the x- and y-coordinates of each sample. Citation frequencies of N descriptor groups from Ultra Flash Profiling are represented as matrix D.

2.4.2 Multiple factor analysis (MFA)

The analysis of projective mapping data is challenging due to the complexity of the data, especially when citation frequencies have to be included into the analysis. The main requirement for multivariate methods is finding a configuration which represents the consensus of the projective maps of all the panellists. To analyse sorting data usually Multidimensional Scaling (MDS) (Lawless et al., 1995) and to a lesser extent DISTATIS (Abdi et al., 2007) and Correspondence Analysis (CA) (Bouteille et al., 2013) are used. For the analysis of data from projective mapping General Procrustes Analysis (GPA) (Risvik et al., 1994) and INDSCAL (Barcenas et al., 2004) have been reported. Multiple Factor Analysis (MFA), which was introduced with the napping approach, is another powerful multivariate method (Pagès, 2005b).

Besides SUMPCA, consensus PCA, STATIS and multiblock correspondence analysis, MFA belongs to the family of multi-block or multi-table PCA methods. All these methods decompose a matrix X, consisting of the submatrices X_k, which are normalised in different manners for each method. MFA can be computed as the PCA of the matrix X, with each submatrix X_k weighted (scaled) by the inverse of its first singular value. The first step of MFA is therefore a PCA for K submatrices X_k with

M rows and N_k columns via their SVD:

$$X_k = U_k S_k (V_k)^T, \tag{2.8}$$

where U_k is an orthogonal $I \times I$-matrix, S_k is a rectangular diagonal $I \times J_k$-matrix with non-negative entries and V_k is an orthogonal $J_k \times J_k$-matrix. The I columns of U_k and the N_k columns of V_k are the left singular vectors and the right singular vectors of X_k. The diagonal entries of S_k are the so-called singular values $\sigma_{1,k} \geq \cdots \geq \sigma_{r,k} > 0$ of X_k, where $r = \min\{I, J_k\}$.

The second step consists of the normalisation of all K submatrices X_k with I rows and J_k columns by the inverse of their first singular values $\sigma_{1,k}$ and subsequent concatenation to the complete final $I \times J$-matrix $\widetilde{Z}$ where $J = \sum J_k$.

$$\widetilde{Z} = \left[\sigma_{1,k}^{-1} X_1 | \sigma_{1,k}^{-1} X_2 | \cdots | \sigma_{1,k}^{-1} X_K\right] \tag{2.9}$$

Each observation can be assigned a mass which reflects its importance. When all observations have the same importance, their masses are all equal to $m_i = \frac{1}{I}$. For reasons of simplicity masses are not taken into account here. A global PCA is finally obtained by Singular Value Decomposition (SVD) of $\widetilde{Z}$:

$$\widetilde{Z} = \widetilde{P}\widetilde{S}(\widetilde{Q})^T, \tag{2.10}$$

In PCA, equation 2.10 is rewritten as

$$\widetilde{Z} = F\widetilde{Q}^T \text{ with } F = \widetilde{P}\widetilde{S} \tag{2.11}$$

where F is a $I \times I$-matrix storing the factor scores (describing the samples/observations) and $\widetilde{Q}$ is $J \times J$-matrix storing the loadings (describing all variable submatrices).

The relationship of the PCAs of each submatrix with the global analysis can be

explored by computing loadings (e.g. correlations) between the components of each submatrix and the components of the global analysis. For more details and examples on the calculation of MFA see Abdi et al. (2013).

CHAPTER III

Development of new approaches for non-targeted GC-MS data analysis

3.1 Introduction

The data generated by hyphenated chromatographic techniques such as GC-MS or LC-MS are especially information rich. Feature extraction such as peak picking or peak integration in single ion chromatograms, total ion chromatograms or deconvoluted signals are the most common approaches to extract information from chromatographic data. The results are in relatively small data tables which are straightforward to analyse (Behrends et al., 2011; Stein, 1999; Aggio et al., 2011; Want and Masson, 2011; Luedemann et al., 2008; Smith et al., 2006; Vestner et al., 2011). Although various peak integration algorithms and software packages have been developed (Dixon et al., 2006; Furbo and Christensen, 2012; Hastings et al., 2002; Vivó-Truyols et al., 2005), automated peak integration remains troublesome due to coelution and potential erroneous peak integration and/or assignment. Time consuming manual correction of the results is often necessary. Moreover, relevant information from the raw data can be lost due to such feature extraction before multivariate data analysis (Skov and Bro, 2005; Ballabio et al., 2008). Deconvoluting chromatographic signals can also be time-consuming in terms of model construction and evaluation of results

(Bro, 1997; Rodríguez et al., 2013; Behrends et al., 2011; Tauler, 1995).

An alternative, more comprehensive approach aimed at the extraction of more information and underlying patterns in the data involves the use of the two dimensional raw data signal (GC-MS chromatogram) of each sample in entirety as a chromatographic fingerprint for modelling. Examples for holistic non-targeted analyses can be found in numerous reports (Ballabio et al., 2008; Sinkov and Harynuk, 2011; Daszykowski et al., 2008; Durante et al., 2011; Cocchi et al., 2008; Durante et al., 2006; Christensen et al., 2005b,a; Christensen and Tomasi, 2007; Silva Ferreira et al., 2014), some of which also include the application of multi-way analysis methods such as Tucker3, PARAFAC and Multi-way Partial Least Squares (N-PLS) to hyphenated chromatographic data. When factor models are used on chromatographic data, challenges are associated with the increased size of data and the handling of shifts and peak shape deformation among chromatograms, which result in distortion of the bilinear/trilinear structure of the data. Several algorithms and software programmes have been developed for peak alignment (Nielsen et al., 1998; Skov et al., 2006; Tomasi et al., 2004; Lange et al., 2007; Sinkov et al., 2011). Depending on the data, shift correction can, however, be difficult and time-consuming.

The above described problems of conventional data analysis approaches to non-targeted GC-MS analysis, in particular challenges with automated peak integration and retention time alignment of chromatograms, were the main motivation for the development of an alternative data analysis approach. The course of the realization and implementation of ideas is described during this chapter. The major consideration to overcome the peak integration issue was the direct modelling of the chromatographic raw data (without feature selection), including a reduction of the data. The main idea to master the distortion of the data structure due to shifting peaks was the use of a mathematical transformation of pieces (segments) of the chromatograms using SSCP matrices. SSCP matrices are positive, squared and symmetric, simi-

lar to variance-covariance matrices (Lay, 2002), which are utilised for instance in PARAFAC2, STATIS and the calculation of R_V-coefficients (Danielsson et al., 2006; Daszykowski et al., 2008; Daszykowski and Walczak, 2011; Stanimirova et al., 2004; Bro et al., 1999). Particularly the indirect fitting algorithm for PARAFAC2 (Harshman, 1972) served as a major inspiration for the development of the new approaches. Moreover, for the sake of simplicity another aim was to use a single model for the entire set of chromatograms of all samples to find systematic differences among samples and to identify important regions of the chromatograms which, if desired, can be further deconvoluted and investigated using e.g. PARAFAC2 or AMDIS. A method using multiple PARAFAC2 models on segmented chromatograms has been reported recently (Amigo et al., 2010a). This approach gives very detailed information on fully decomposed mass spectra and peak profiles, which are finally summarized using PCA. The here described new approach can be considered as a 'segment pre-selection tool' for subsequent deconvolution of only important chromatogram segments. By this means a significant amount of time used for the deconvolution of chromatogram segments (e.g. construction and evaluation of PARAFAC2 models) can be save.

This chapter gives an overview on the algorithms of the new data analysis approaches, including the theoretical background on mathematical transformations such as the calculation of SSCP matrices and SVD. The approaches are explained and tested on an artificial, well defined GC-MS data set with and without peak shifts. Moreover, the limitations of the established methods such as PCA and Tucker3 on the artificial GC-MS raw data in terms of variable size and peak shifts are discussed. After the theoretical discussion, the approaches are demonstrated on a real GC-MS dataset of experimental wines and results are confirmed using a reference method including PARAFAC2 deconvolution and peak integration of deconvoluted peak profiles of the entire segmented chromatograms with subsequent PCA on the obtained peak table.

3.2 Defined, artificial GC-MS data set

To demonstrate and verify the developed algorithms a defined, artificial GC-MS data set was created using an in-house developed MATLAB script. The data set consists of 20 chromatograms, each containing 9 to 10 gaussian peaks with different mass spectra (35 u to 318 u) and different degrees of overlapping. The whole chromatogram can be divided into five segments. Segment one contains two peaks which perfectly overlap. Peaks three and four partially coelute in segment two, which is also the case for the peaks five, six and seven in segment three. Peak eight is in segment four and the last segment contains the last two peaks nine and ten, which also partially coelute (Figure 3.1 and 3.2). Peak sizes of three peaks vary among chromatograms as indicated in Table 3.1, consequently samples can be divided into four groups. Moreover, a small random variation was added to all peak sizes to simulate a natural deviation of measurements. To simulate baseline noise a random normal distributed noise was added to the whole data set. Each chromatogram can be considered as a matrix of dimensions *1100 scans* $\times$ *283 masses*, thus the entire data set can be considered as a three-way array ($i \times j \times k$), with the dimensions *20 samples* $\times$ *1100 scans* $\times$ *283 masses*.

segment	peak no.	size difference	sample no.
1	2	only present in	14 & 15
2	4	0.7$\times$ higher in	1 to 5
5	9	3$\times$ higher in	1 to 10

Table 3.1: Differing peaks (No. 2, 4 and 9) among samples in the defined, artificial GC-MS data set. All other peaks are of the same size in all samples.

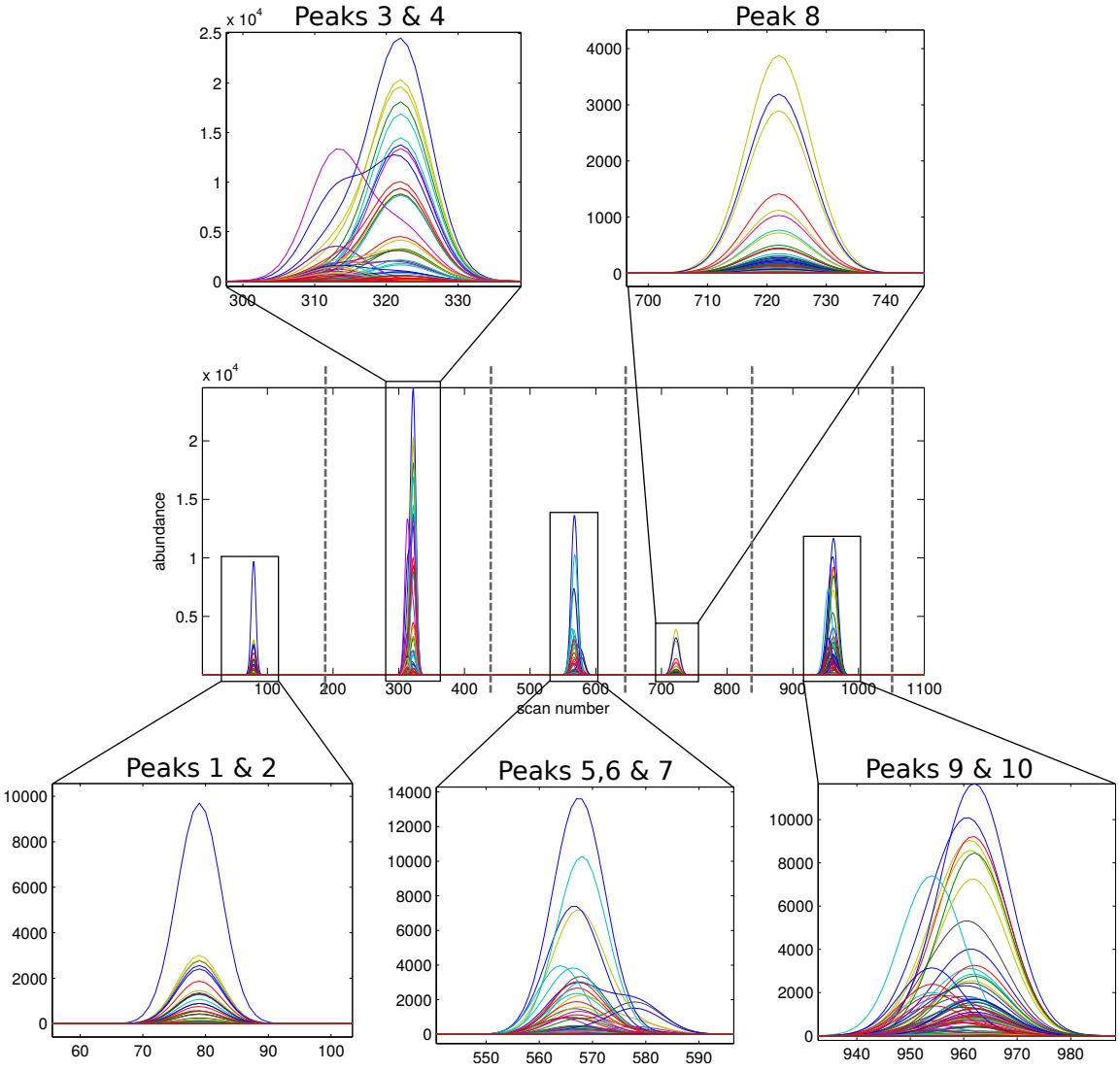


Figure 3.1: Overlay of all mass channels of one sample (sample no. 14) of the artificial GC-MS data set. Dotted lines show the segmentation of the chromatograms.

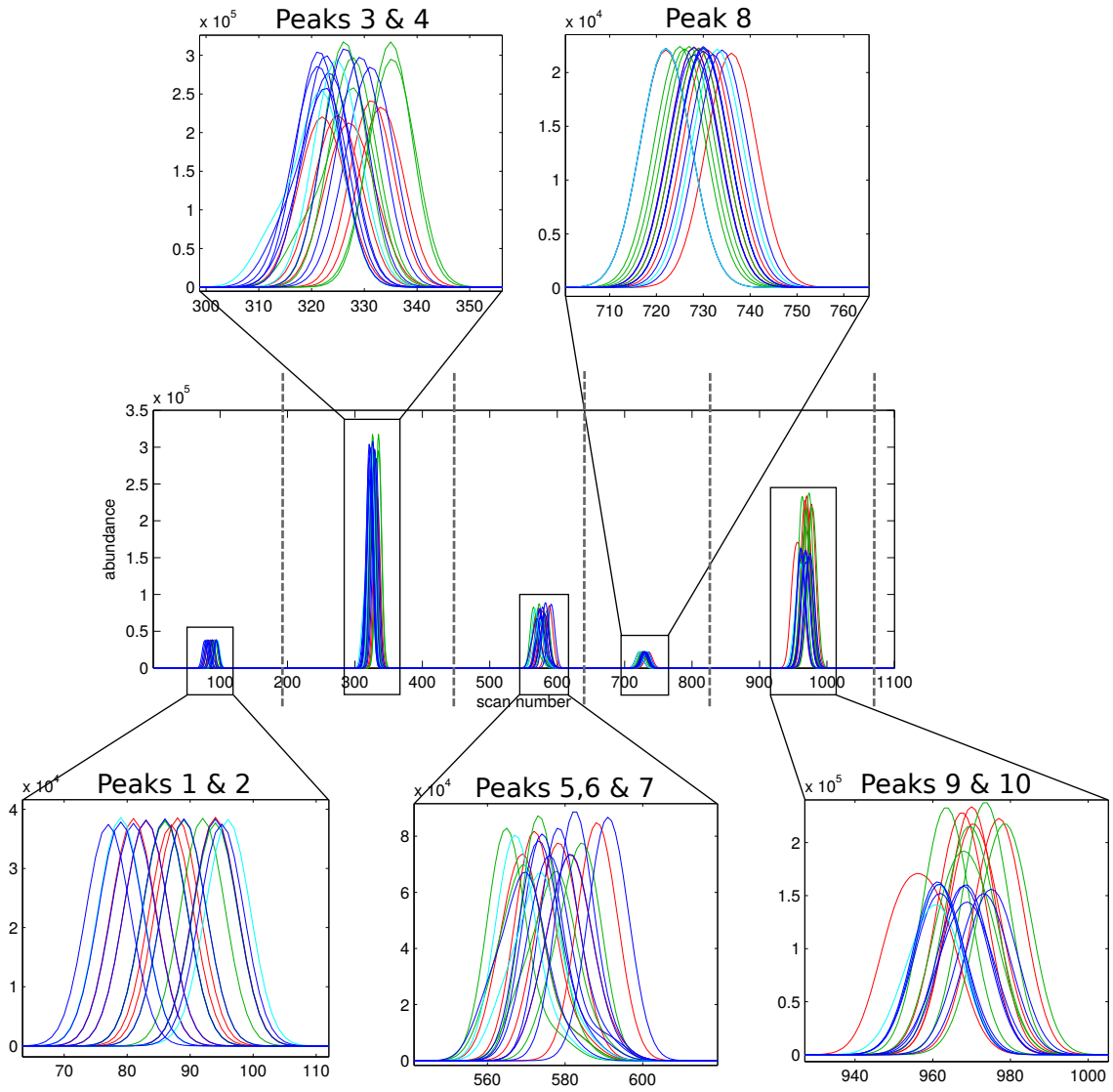


Figure 3.2: Overlay of TICs of all samples of the artificial GC-MS data set with introduced shift. Dotted lines show the segmentation of the chromatograms.

3.3 Limitations of PCA and Tucker3 on chromatographic raw data

Feature selection such as automated integration of peaks is not needed, when multivariate models are used directly on chromatographic raw data. A large number of variables and shifting retention time profiles pose problems for multivariate models in terms of the distortion of the bilinear/trilinear structure of the data and in terms of reasonable stability and reliability of multivariate analyses, respectively.

3.3.1 Artificial GC-MS data without peak shifts

To demonstrate the above mentioned issues PCA was applied to the TIC of all samples of the artificial GC-MS data set without peak shifts as well as on the entire unfolded three-way array which was rearrange in a way that the mass spectral dimension was eliminated as indicated in Figure 3.3. Furthermore, taking the multiway nature of the artificial data set into account Tucker3 was used to decompose the three-way array. For preprocessing auto-scaling and mean-centering was used.

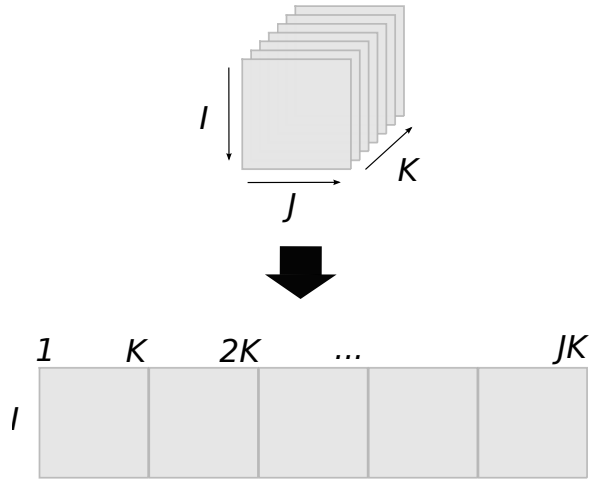


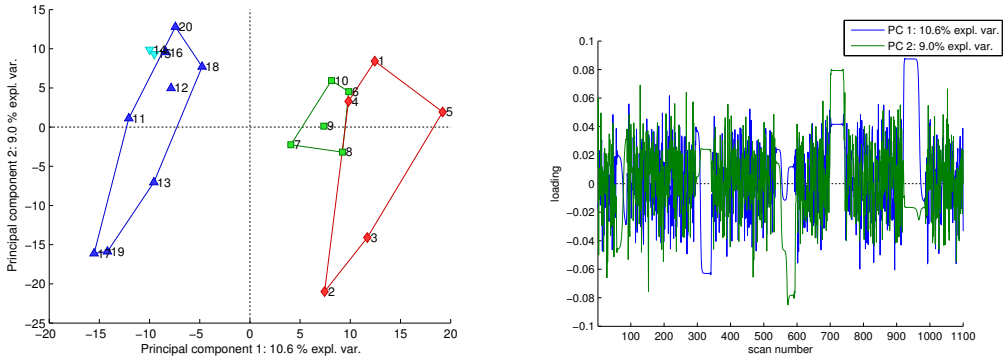
Figure 3.3: Unfolding of the three-way array ($i \times j \times k$), where i is the number of samples, j is the elution profile (number of scans) and k is the number of mass channels, into a new matrix ($i \times jk$).

Figure 3.4 shows the scores and loading plots of a PCA on the autoscaled TICs

(without peak shift). It can be observed that the first ten and the last ten samples are differentiated on Principal Component (PC) one. This separation is caused by the large difference of peak nine among these samples as can be observed in the loadings plot. However without the *a priori* knowledge of the artificial data set it would be very difficult to draw this conclusion as peaks nine and ten in the chromatogram are partially coeluted. All variables get the same weight in PCA, when auto-scaling is used. The here presented chromatographic data set consists of 20 samples and 1100 variables, of which many contribute to baseline noise due to the sparse nature of chromatographic data. It is evident that after auto-scaling the baseline noise is extremely up weighted as can be seen in the loading plot in Figure 3.4(b). This also explains the low proportion of explained variance by PC one of 10.6 % and by PC two of 9.0 %. The scores and loadings plots of the PCA on the unscaled data, which was only mean-centered, is shown in Figure 3.5. A very clear grouping of the samples one to five, six to ten and eleven to 20 can be observed on PC one (81.3 % explained variance) and on PC two (18.6 % explained variance), respectively. All other PCs does not explain any structural information. Again, with the *a priori* knowledge of the artificial data set it is clear that the variables showing high loadings on the corresponding principal components (Figure 3.5(b)) can be assigned to the peaks four and nine.

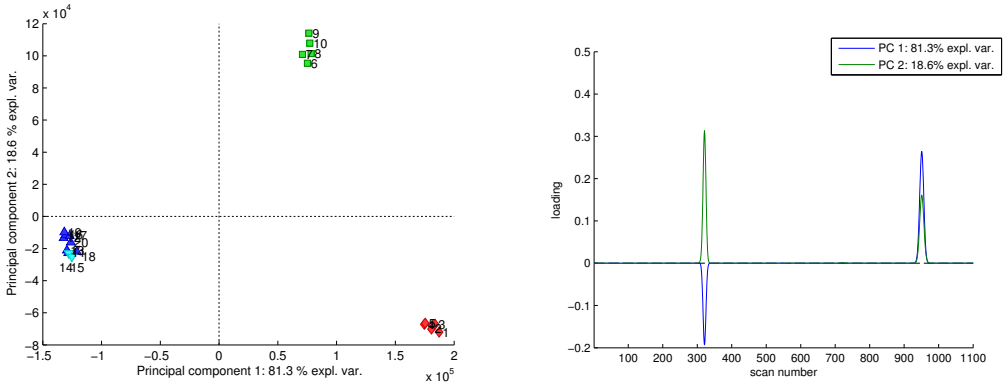
If variables are not scaled to unit variance prior to PCA, all variables with the highest variance or standard deviation, respectively, will have the biggest influence on the model. In other words larger variables will evidently have a larger influence on the model than smaller variables. Although samples grouped very well together in the here presented example, it might not be a good idea to use only mean-centering in a real world situation, as small, but important variables (or peaks), could easily be missed. Moreover, in the here presented example the samples 14 and 15 which contrary to the other samples contain peak number two can not be separated. As the

TICs are the sum of all mass channels, the information on the relative small peak number 2 in the samples 14 and 15 is simply lost during the summation of all masses.



(a) Scores: PC 1 vs PC 2 (b) Loadings: PC 1 vs PC 2

Figure 3.4: Scores and loadings plot of the first two principal components of the PCA (auto-scaled) on the TICs of the artificial data set (without peak shifts). Samples are coloured according to Table 3.1.

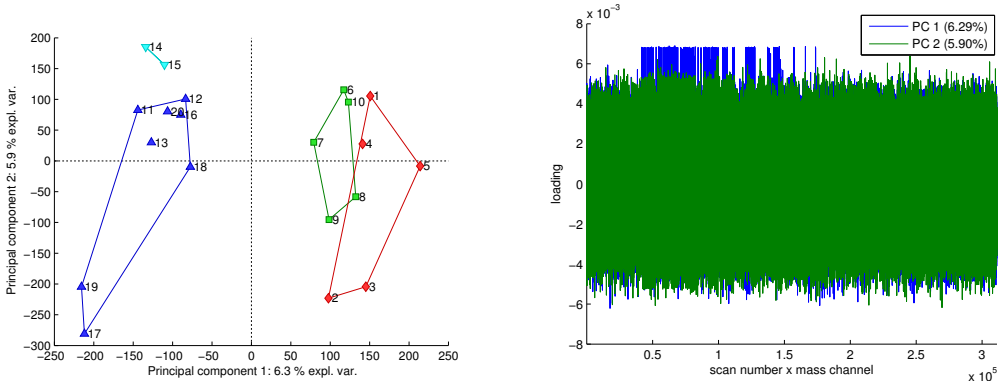


(a) Scores: PC 1 vs. PC 2 (b) Loadings: PC 1 vs. PC 2

Figure 3.5: Scores and loadings plot the first two principal components of the PCA (mean-centered) on the TICs of the artificial data set (without peak shifts. Samples are coloured according to Table 3.1.

It is possible to prevent the loss of the information of the mass dimension by unfolding the three-way array prior to PCA as shown in Figure 3.3. The unfolding of the array ($i \times j \times k$), with i number of samples, j scans (the elution profile) and k mass channels results in a new matrix ($i \times jk$). In the here presented example this matrix is of size 20×311300 ($20 \times (1100 \times 283)$). In fact this matrix is much

bigger than the matrix of TICs, which worsen the issue of an excessive number of variables for PCA modelling in terms of reasonable stability and reliability of the model. The Figures 3.6 and 3.7 show the scores and loadings plots of the first two components of the PCAs on the unfolded three-way array with auto-scaling and with mean-centering only, respectively. The first two principal components of the PCA on the auto-scaled data (Figure 3.6) explain a very low amount of 6.3 % and 5.9 % of the total variance in the data set. The first two principal components of the PCA on the mean-centered data (Figure 3.7) explain 84.2 % and 15.7 % of the variance in the data. All other PCs do not explain any structural information. The results are very similar to the above discussed PCAs on the TICs. Although, the mass dimension remained intact when the three-way array was unfolded, the differences of the samples 14 and 15 which are the only samples containing the relative small peak number two are not well reproduced in the two PCAs. Without scaling to unit variance all variables with small variances have very little influence on the PCA. Scaling to unit variance, however, extremely up weight noise, as stated above already.

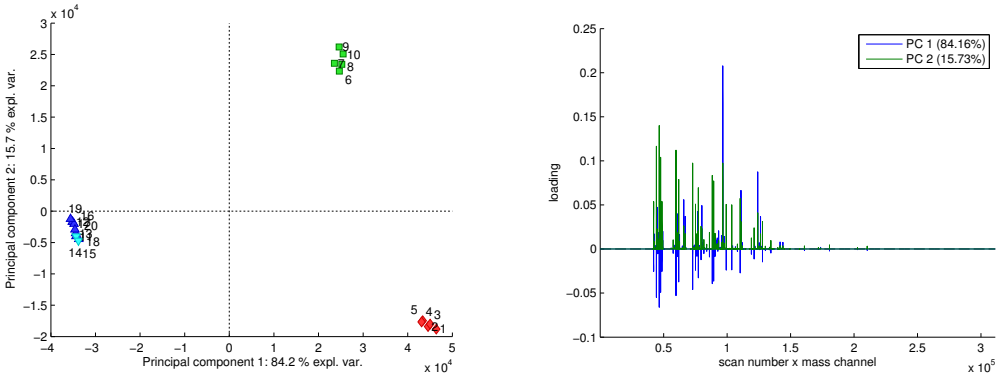


(a) Scores: PC 1 vs. PC 2

(b) Loadings: PC 1 vs. PC 2

Figure 3.6: Scores and loadings plots of the first two principal components of the PCA (autoscaled) on the unfolded three-way array (Figure 3.3) of the artificial data set (without peak shifts). Samples are coloured according to Table 3.1.

The interpretation of the loadings of the PCAs on the unfolded three-way array is very difficult (Figures 3.6(b) and 3.7(b)). Unfolding mixes up variables in the



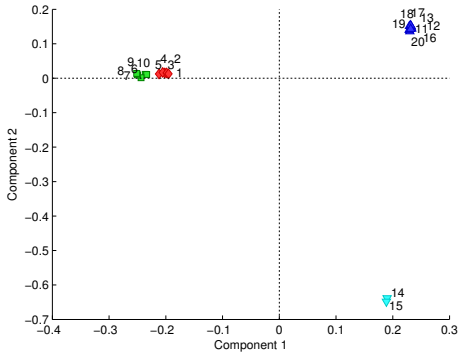
(a) Scores: PC 1 vs. PC 2 (b) Loadings: PC 1 vs. PC 2

Figure 3.7: Scores and loadings plots of the first two principal components of the PCA (mean-centered) on the unfolded three-way array (Figure 3.3) of the artificial data set (without peak shifts). Samples are coloured according to Table 3.1.

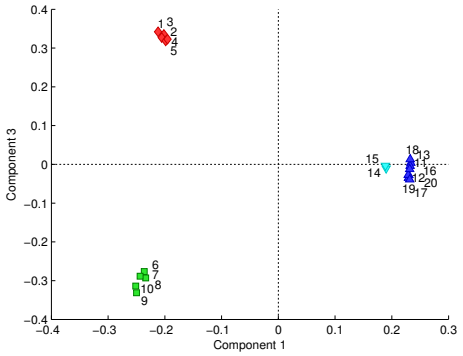
unfolded mode, so that the effect of one variable is associated with more than one element of a loading vector. A two-way PCA model can therefore be considered to be less simple compared to a multi-way model such as Tucker3. With orthogonal factors Tucker3 is also known as multi-way PCA. The two component PCA model on the 20×311300 unfolded three-way array consists of 622640 parameters ($2 \times 20 + 2 \times 311300$), whereas a [2 2 2] component TUCKER model of the $20 \times 1100 \times 283$ three-way array consists of 2814 parameters ($2 \times 20 + 2 \times 1100 + 2 \times 283$). This example shows that Tucker3 can be considered to be the simpler model in a multi-way context. Moreover, it is apparent that this multi-way model is much easier to interpret.

A [3 3 3]-Tucker3 model on the artificial GC-MS dataset which was mean-centered across the first mode (samples) and scaled to unit variance within the third mode (mass channels) was constructed. In order to simplify the interpretation of the model the initial Tucker3 core was rotated to optimal diagonality (Table 3.2). The loadings of all modes of the Tucker3 model are shown in Figure 3.8. The first component explains the difference between the samples one to ten and the samples eleven to twenty (Figure 3.8(a)). This caused by peak nine which negatively correlates with component one

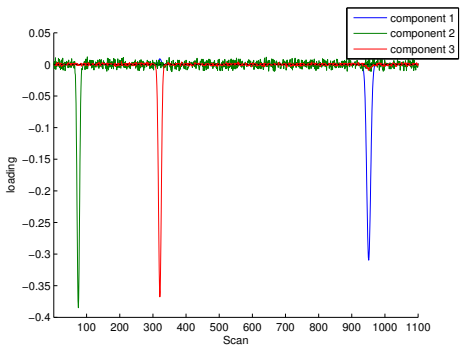
(Figure 3.8(c)). The difference of the samples 14 and 15 to all other samples is represented by component two, which correlates negatively with peak number two (see Figure 3.8(c)). Component three represents the difference between the samples one to five and six to ten (Figure 3.8(b)). These samples are associated with peak number four, which negatively correlates with this component (Figure 3.8(c)). The loadings of the third mode which indicate the importance of the mass channels to each of the components are shown in Figure 3.8(d). These results show that the Tucker3 model is more appropriate than two-way PCA on the unfolded array to extract all relevant structural information out of the artificial GC-MS data set. Tucker3 is advantageous over the two-way approach, because the nature of the model corresponds to the nature of the data as has been discussed above (multi-way models for multi-way data).



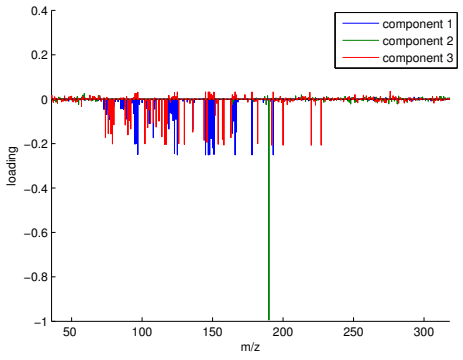
(a) Mode 1: comp. 1 vs. comp. 2



(b) Mode 1: comp. 1 vs. comp. 3



(c) Mode 2: comp. 1 to comp. 3



(d) Mode 3: comp. 1 to comp. 3

Figure 3.8: Loadings of modes one to three of the Tucker3 model on the three-way array of the artificial GC-MS dataset (without peak shifts). Samples are coloured according to Table 3.1.

	Index to elements	core entry	Explained variation of the core
1	(1, 1, 1)	-267.2	74.0 %
2	(2, 2, 2)	-115.7	13.7 %
3	(3, 3, 3)	-84.8	7.5 %
4	(1, 3, 3)	53.7	2.9 %
5	(1, 2, 2)	33.4	1.2 %
6	(3, 1, 1)	17.0	0.3 %
7	(2, 1, 1)	14.4	0.2 %
8	(3, 3, 1)	-10.4	0.1 %

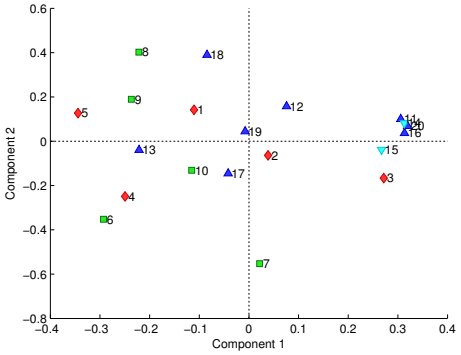
Table 3.2: Eight largest core entries and their corresponding explained variation (sum of squares) of the [3 3 3]-TUCKER model on the three-way array of the artificial GC-MS data set (sorted in descending order).

3.3.2 Artificial GC-MS data with peak shifts

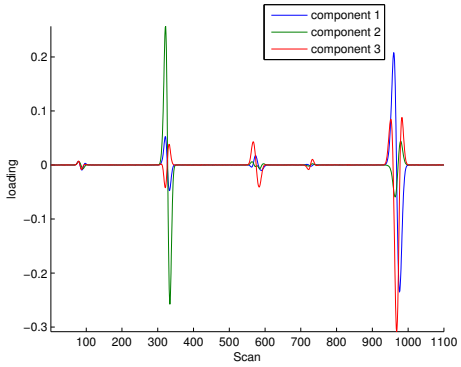
In the previous section all models have been tested on a data set which did not contain any shifts in the retention profile among samples. Nevertheless, experimental chromatographic data most often contain shifts among samples. Before factor models can be applied directly on real chromatographic data peak alignment is inevitable. Figure 3.9 shows the loadings of a [3 3 3]-Tucker3 model on the artificial GC-MS data with introduced non-linear peak shifts for every peak. An overlay of all TICs of this shifted data set is displayed in Figure 3.2. The loadings of the sample mode (mode one) of the [3 3 3]-Tucker3 model, presented in Figure 3.9(a) (only component one vs. component two), show that all samples randomly scatter and no structural information on the different groups of samples is obtained. In Figure 3.9(b), which shows the loadings of the second mode (elution profile) on component one to three, typical patterns for loadings of shifted peaks which look similar to the first derivative of a peak can be observed.

The possibility of the usage of multi-way models such as Tucker3 to decompose multi-way chromatographic data such as GC-MS chromatograms from multiple samples has been demonstrated in the previous chapter. Feature selection such as auto-

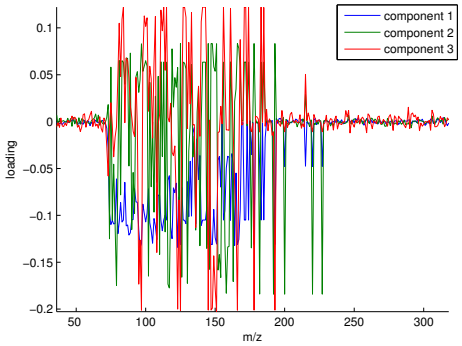
mated peak integration which, dependent on the data, can be troublesome can so be avoided. Peak shifts and a serious problem when factor models are directly applied to chromatographic raw data as has been demonstrated in the last example (Figure 3.9). All in all, this issue shows the necessity for further development of data analysis approaches which take shifting peaks into account.



(a) Mode 1: comp. 1 vs. comp. 2



(b) Mode 2: comp. 1 to comp. 3



(c) Mode 3: comp. 1 to comp. 3

Figure 3.9: Loadings of modes one to three of the Tucker3 model on the three-way array of the artificial GC-MS dataset with shifted peaks. Samples are coloured according to Table 3.1.

3.4 Approach 1: 'chromatogram segmentation, SSCP matrices and PARAFAC'

The in this and the following section discussed development of two new approaches to GC-MS data analysis were primarily inspired by the indirect fitting algorithm for PARAFAC2 (Harshman, 1972), in which SSCP matrices are used to compensate for the distortion of the trilinearity of three-way data. The aim was the usage of a single model for the whole chromatograms of all samples to obtain information on systematic differences among samples. Out of the principal idea of the indirect fitting algorithm for PARAFAC2 a new idea was developed that makes the modelling of the entire chromatograms of all samples possible by implementing segmentation and mathematical transformation of chromatogram segments of each sample into SSCP matrices. In this manner the new approaches cope without peak integration and peak alignment.

In the following the basic ideas and the development of a first approach are discussed. This first approach includes segmentation of chromatograms, mathematical transformation of chromatogram segments using SSCP matrices and PARAFAC modelling of the obtained three-way array of the transformed chromatographic raw data.

3.4.1 Theoretical background

Using basic matrix algebra a SSCP matrix XX^T is obtained by multiplication of a matrix X with its transpose, as displayed in Equation 3.1.

$$XX^T = \begin{bmatrix} \sum_{j=1}^{C} x_{1j}^2 & \sum_{j=1}^{C} x_{1j}x_{2j} & \cdots & \sum_{j=1}^{C} x_{1j}x_{Rj} \\ \sum_{j=1}^{C} x_{2j}x_{1j} & \sum_{j=1}^{C} x_{2j}^2 & \cdots & \sum_{j=1}^{C} x_{2j}x_{Rj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{C} x_{Rj}x_{1j} & \sum_{j=1}^{C} x_{Rj}x_{2j} & \cdots & \sum_{j=1}^{C} x_{Rj}^2 \end{bmatrix}, \tag{3.1}$$

where X is a $R \times C$-matrix of elements x_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$. The matrix product XX^T is the $R \times R$ matrix of Sums of Squares and Cross Products (SSCP matrix). In Detail, the diagonal of XX^T includes the sums of squares with respect to a given row i of X, namely $\sum_{j=1}^{C} x_{ij}^2$. Moreover, all off-diagonal elements represent cross products between two different rows i, k of X, in particular $\sum_{j=1}^{C} x_{ij}x_{kj}$ for $i \neq k$. Consequently, the sums of squares are a measure of variation within a row, whereas the cross products are a measure of covariation between two rows. Note the similarity to the variance-covariance matrix: diagonal elements of the variance-covariance matrix are variances and all off-diagonal elements are covariances[1]. The terms variation and variance as well as covariation and covariance can for the sake of simplicity be replaced in the following (although not strictly mathematically true).

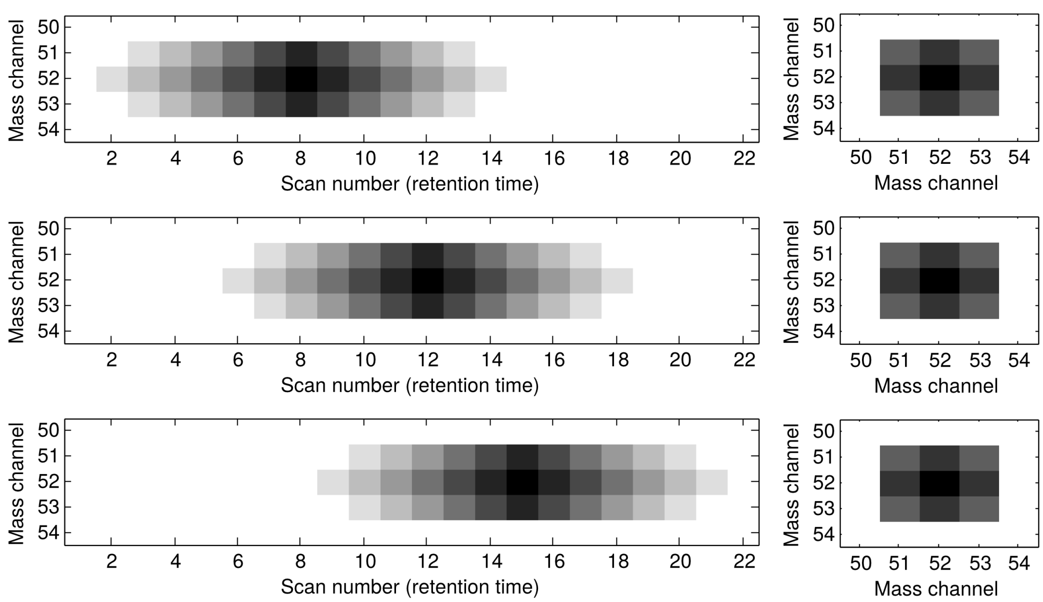
PARAFAC2 is a powerful tool for the deconvolution of small chromatogram segments (Bro et al., 1999; Amigo et al., 2010a, 2008; Johnsen et al., 2014). The approach presented here is mainly inspired by the idea of the indirect fitting algorithm of the PARAFAC2 model, which instead of modelling an array consisting of the matrices X^i (*spectral profile* $\times$ *elution profile for I samples*) directly considers a model of an array consisting of the SSCP matrices $X^i(X^i)^T$ (Bro, 1998b; Harshman, 1972). In this manner, PARAFAC2 is suitable for deconvoluting chromatographic peaks shifting along the retention axis among samples. A disadvantage of PARAFAC2 is that for each segment of the chromatogram a single model has to be constructed and evaluated. Figure 3.10 shows an visualised example of three identical but shifted two dimensional GC-MS peaks (simulated data), represented as the matrices X, Y and Z, and their SSCP matrices XX^T, YY^T and ZZ^T. The table in 3.10(a) elucidates that the three SSCP matrices are constant.

The utilisation of SSCP matrices as a preprocessing step for multivariate mod-

[1]If the means of the columns of X are zero, the summed cross-products for two variables will be proportional to their covariances (covariance matrix). If, in addition, the variances of the columns of X are unity, then the cross-products for two variables will be equal to the correlation coefficient for those two variables (correlation matrix).

	Raw signal data points																							SSCP matrices				
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	0	0	0	0	0		0	**584**	**728**	**584**	0
X	0	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	0	0	0	0	XX^T	0	**728**	**924**	**728**	0
	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	0	0	0	0	0		0	**584**	**728**	**584**	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
	0	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	0		0	**584**	**728**	**584**	0
Y	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	YY^T	0	**728**	**924**	**728**	0
	0	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0	0	0	0		0	**584**	**728**	**584**	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
	0	0	0	0	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0		0	**584**	**728**	**584**	0
Z	0	0	0	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**12**	**10**	**8**	**6**	**4**	**2**	0	ZZ^T	0	**728**	**924**	**728**	0
	0	0	0	0	0	0	0	0	0	**2**	**4**	**6**	**8**	**10**	**12**	**10**	**8**	**6**	**4**	**2**	0	0		0	**584**	**728**	**584**	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0

(a) Numeric representation



(b) Visual representation

Figure 3.10: Three simulated two dimensional GC-MS peaks consisting of 22 scans (retention time) and 5 mass channels, represented as the matrices X, Y and Z, and their SSCP matrices XX^T, YY^T and ZZ^T (modified from van Mispelaar et al. (2003)).

elling of whole chromatograms has been reported before (Daszykowski et al., 2008; Daszykowski and Walczak, 2011). If entire two dimensional chromatograms are used for the construction of SSCP matrices, information on the retention time of compounds is lost, complicating the identification of peaks contributing to the differentiation among samples.

However, by dividing all chromatograms along the retention axis into segments containing a small number of peaks and subsequent construction of SSCP matrices for each segment, information on the location of peaks in the chromatogram contributing to the differentiation of samples can be preserved. In the here presented new approach 1, the SSCP matrices for each segment and each sample have dimensions *number of mass channels* $\times$ *number of mass channels* and contain information on the variation of each mass channel and covariation between all mass channels in each segment for the corresponding sample. For each segment the constructed SSCP matrices of all samples are vectorized and compiled into a new matrix. This step results in a compilation matrix for each segment with the dimensions *number of samples* $\times$ *[(number of mass channels* + 1*)* $\cdot$ *number of mass channels* / 2*]*.

These compilation matrices are then also transformed into SSCP matrices with the dimensions of *number of samples* $\times$ *number of samples*, which contain information about the variation of the content of the compilation matrix for each sample and the covariation of the content of the compilation matrix between all samples in each segment. These SSCP matrices are finally compiled in a three-way array with the dimension *(number of samples* $\times$ *number of samples)* $\times$ *number of segments*.

The whole procedure is summarized in matrix notation in the following. Each two dimensional chromatogram (sample) is characterized by M mass channels and N scan points. N is divided into K segments, that is $N = \sum_{k=1}^{K} N_k$, where N_k describes the number of scans in the k-th segment. In particular, altogether we have I samples.

First, we define an $I \times K$-matrix X by

$$X = (X^{ik})_{\substack{i=1,\ldots,I \\ k=1,\ldots,K}} = \begin{bmatrix} X_{11} & \cdots & X_{1K} \\ \vdots & \ddots & \vdots \\ X_{I1} & \cdots & X_{IK} \end{bmatrix}, \tag{3.2}$$

where X^{ik} is a $M \times N_K$-matrix containing the data of the i-th sample and k-th segment, that is

$$X^{ik} = (x^{ik}_{mn})_{\substack{m=1,\ldots,M \\ n=1,\ldots,N_k}} = \begin{bmatrix} x^{ik}_{11} & \cdots & x^{ik}_{1N_k} \\ \vdots & \ddots & \vdots \\ x^{ik}_{M1} & \cdots & x^{ik}_{MN_k} \end{bmatrix}. \tag{3.3}$$

The SSCP matrix $A^{ik} = X^{ik}(X^{ik})^T$ containing information on the variation and covariation between all mass channels of the i-th sample and k-th segment is defined by

$$A^{ik} = (a^{ik}_{rt})_{r,t=1,\ldots,M} \tag{3.4}$$

$$\text{with } a^{ik}_{rt} = \sum_{s=1}^{N_k} x^{ik}_{rs} x^{ik}_{st} \qquad \forall r, t = 1, \ldots, M \tag{3.5}$$

$$\text{and } \dim(A^{ik}) = M \times M, \tag{3.6}$$

for all $i = 1, \ldots, I$ and $k = 1, \ldots, K$.

Subsequently only the upper triangular part of the symetric SSCP matrix A^{ik} is vectorised (unfolded) and concatenated into a new matrix Y^k (compilation matrices).

The vectorisation $\mathrm{vec}(A^{ik})$ of the upper triangular of A^{ik} is defined by[2]

$$\mathrm{vec}(A^{ik}) = \alpha_1^{ik} \frown \alpha_2^{ik} \frown \cdots \frown \alpha_M^{ik}, \tag{3.7}$$

where

$$\alpha_l^{ik} = (a_{l,l}^{ik}, a_{l,(l+1)}^{ik}, \ldots, a_{l,M}^{ik}) \qquad \forall l = 1, \ldots, M, \tag{3.8}$$

for all $i = 1, \ldots, I$ and $k = 1, \ldots, K$.

Consequently, the vectorisation $\mathrm{vec}(A^{ik})$ has $J = \sum_{l=1}^{M} l = \frac{M(M+1)}{2}$ components. The $I \times J$-matrix Y^k is constructed by the above row vectors $\mathrm{vec}(A^{1k})$, ..., $\mathrm{vec}(A^{Ik})$ as follows:

$$Y^k = \begin{bmatrix} \mathrm{vec}(A^{1k}) \\ \vdots \\ \mathrm{vec}(A^{Ik}) \end{bmatrix}, \tag{3.9}$$

for all $k = 1, \ldots, K$.

In the end, we form SSCP matrices $Z^k = Y^k (Y^k)^T$, which contain information on the variation and covariation between all samples in the k-th segment with regard to the variation and the covariation between all mass channels of the i-th sample and k-th segment,

$$Z^k = (Z_{rs}^k)_{r,s=1,\ldots,I} \tag{3.10}$$

$$\text{with } Z_{rs}^k = vec(A^{rk}) \cdot (vec(A^{sk}))^T \qquad \forall r, s = 1, \ldots, I, \tag{3.11}$$

for all $k = 1, \ldots, K$. Finally, the matrices Z^k are rearranged into the $(I \times I) \times K$-array

[2]The concatenation $\frown$ of two arbitrary row vectors $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ is defined as:

$$x \frown y = (x_1, \ldots, x_n, y_1, \ldots, y_n).$$

$\underline{Z}$:

$$\underline{Z} = \left[Z^1 \quad \cdots \quad Z^K \right]. \tag{3.12}$$

Prior to multi-way analysis the three-way array $\underline{Z}$ is mean centered across the first and second mode and scaled to unit variance within the third mode. The term mode refers here to the dimension of the array.

3.4.2 Application of approach 1 to the artificial GC-MS data set

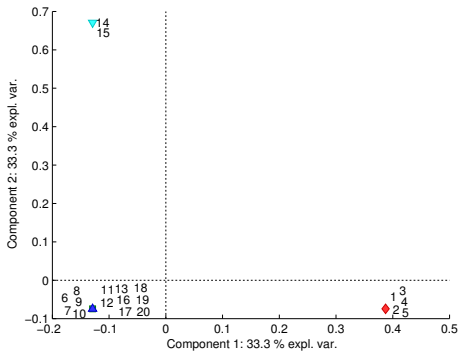
The artificial GC-MS data set was analysed using the new approach to show its validity. To prove theoretical considerations the new approach was first tested on the artificial GC-MS data set without noise and without peak shift. Subsequently, the new approach was tested on the artificial GC-MS data set with noise and non-linear peak shifts to show that the new algorithm can compensate peak shifts.

In the artificial GC-MS data set each of the three differences among groups of samples (see Table 3.1) is caused by varying peak sizes in different segments. After segmentation and mathematical transformation the resulting three-way array contains information on the covariation among samples in terms of differences in their mass traces in each segment. The decomposition of this array using PARAFAC is therefore expected to give one component to explain each of the three differences among the four groups of samples. Noise was excluded from the artificial data set, as it is a source of random variation.

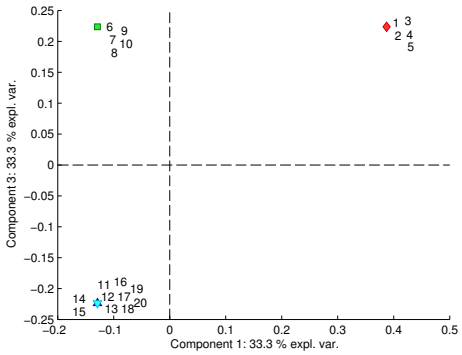
In fact, after applying approach 1 a three component PARAFAC model fully decomposes the segmented and transformed three-way array. The proper number of components was determined by evaluating residuals, core consistency, iterations until convergence, and by assessing the interpretability of the solution. As no noise was introduced to the artificial GC-MS data set 100 % variation is explained, evenly distributed over the three components. The loadings of the first and third mode

(sample and segment mode) are shown in Figure 3.11. Note that due to the calculation of SSCP matrices included in the mathematical transformation modes one and two are identical. Component one explains the differences between samples one to five and the other samples, which is caused by peak four in segment two as indicated by the loadings of mode three of this component. PARAFAC component two reflects the differences of samples 14 and 15, which are the only samples that contain peak number two, in segment one. Finally, the differences between the samples one to ten and eleven to 20 are shown by component three. Here segment five containing peak nine is responsible for this separation. These results are in accordance to the results obtained by Tucker3 analysis of the non-shifted GC-MS raw chromatograms of the artificial data set (Figure 3.8).

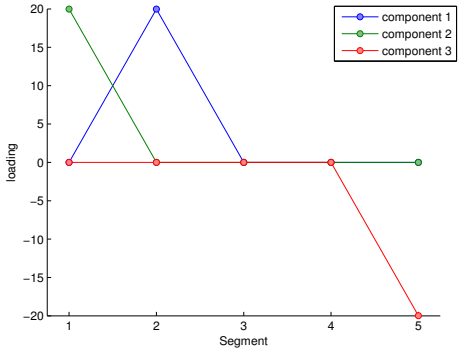
To prove the applicability of the new algorithm to shifted chromatograms the artificial GC-MS data set with introduced peak shifts and random noise (Figure 3.2), which was also used to demonstrate the limitations of the Tucker3 model on the raw chromatograms, was analysed. After segmentation and mathematical transformation a four component PARAFAC model explaining 83.8 % of the total variation in the data was obtained. The proper number of components was determined by evaluating residuals, core consistency, iterations until convergence, and by assessing the interpretability of the solution. Component one explaining 68.6 % of the total variation in the data separates samples one to ten from samples eleven to 20 (Figure 3.12(a)). Segment five, which contains peak number 9 shows, high loadings on this component (Figure 3.12(d)). The samples one to five differ from the other samples on component two, which explains 9.5 % of the total variation. The loadings of the segment mode (mode three) reveal that segment two containing peak four is responsible for this difference. The samples 14 and 15, which as only samples contain peak number 2, are differentiated from the other samples on component three explaining 5.5 % of the total variation (Figure 3.12(b)). Here segment one shows high loadings on this



(a) Mode 1: comp. 1 vs. comp. 2

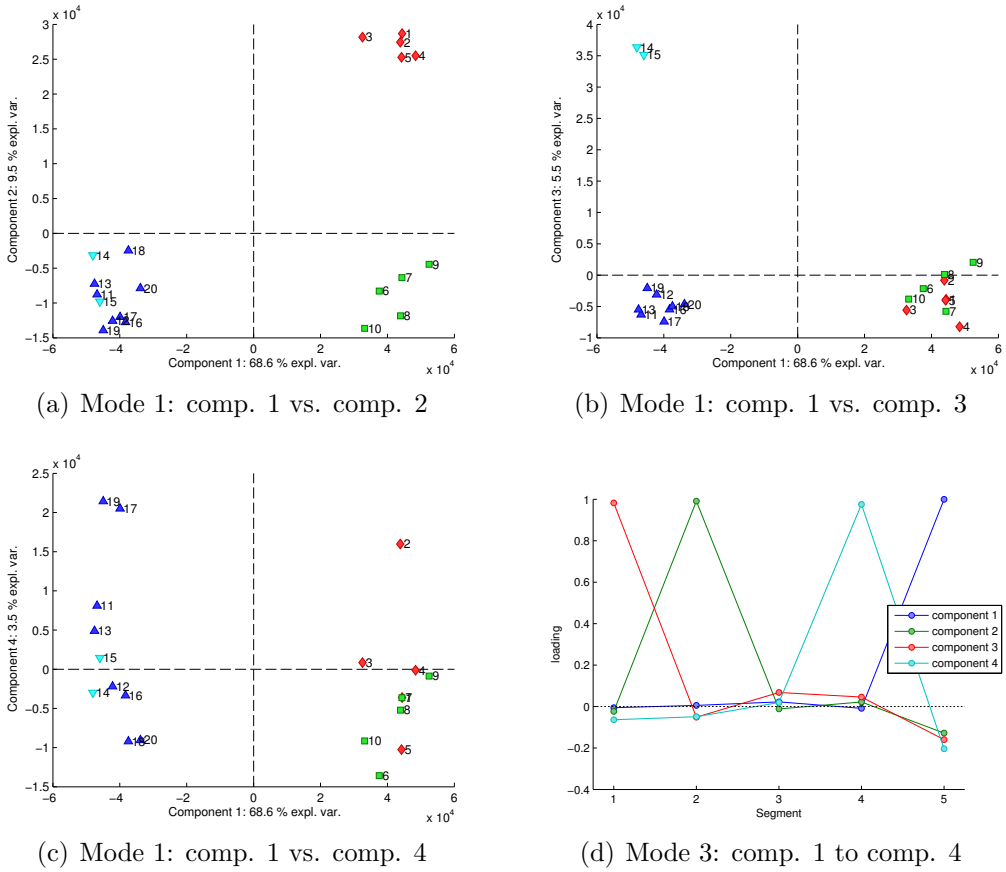


(b) Mode 1: comp. 1 vs. comp. 3



(c) Mode 3: comp. 1 to comp. 3

Figure 3.11: Loadings of the modes one and three of the PARAFAC model on the three-way array of the segmented and mathematically transformed artificial GC-MS dataset without noise and without shifted peaks. Note that mode one and two are identical. Samples are coloured according to Table 3.1.



(a) Mode 1: comp. 1 vs. comp. 2

(b) Mode 1: comp. 1 vs. comp. 3

(c) Mode 1: comp. 1 vs. comp. 4

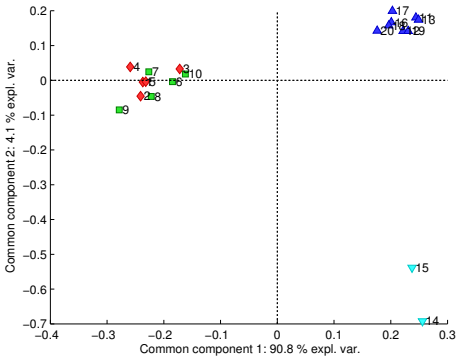
(d) Mode 3: comp. 1 to comp. 4

Figure 3.12: Loadings of the modes one and three of the PARAFAC model on the three-way array of the segmented and mathematically transformed artificial GC-MS dataset with shifted peaks and noise. Note that mode one and two are identical. Samples are coloured according to Table 3.1.

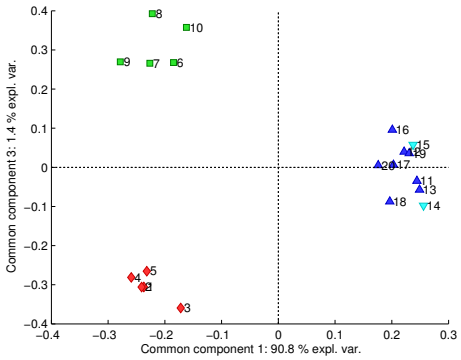
component. Furthermore, component four explaining 3.5 % variation reflected unsystematic variation in the data (Figure 3.12(c)), which is related to noise. Note that a PARAFAC model on the shifted artificial GC-MS data set which does not contain noise results in a three component model (model not shown). Overall, the same structural information on the differences among samples could be extracted from the artificial GC-MS data set with and without peak shifts using the developed approach.

The three-way data array which is obtained after the segmentation and mathematical transformation can also be seen as a 'stack' of matrices. It seems therefore reasonable to evaluate different multi-block methods for the analysis of this data

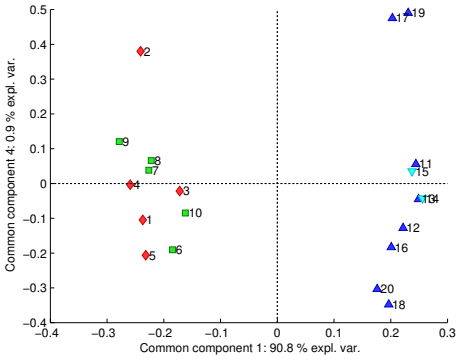
type besides multi-way methods. Different multi-block methods have been applied to the three-way array, in such a manner such that each slab of the array corresponds to a segment. The following methods were tested: PCA on concatenated matrices, MFA (Escofier and Pagès, 2008), Common Components and Specific Weights Analysis (CCSWA)(Mazerolles et al., 2006), analysis of co-inerita with common components (Chessel and Hanafi, 1996) and STATIS (Stanimirova et al., 2004) using the SAISIR toolbox for MATLAB (Cordella and Bertrand, 2014) kindly and freely available on www.chimiometrie.fr (July 2014). From the tested models only CCSWA gave interpretable results which are shown in Figure 3.13. A CCSWA model with 4 components revealed the structural information in the data (Figure 3.12) comparable to the results from PARAFAC (Figure 3.12). Common component one (90.8 % explained variance) separates the samples one to ten and eleven to 20, while segment five has the strongest influence on this component. Common component two (4.1 % explained variance) explains differences between the samples 14 and 15 and the other samples (Figure 3.13(a)). Segment two shows the highest weight on this component. The differences among the samples one to five from the other samples are explained by common component three (Figure 3.13(b)), on which segment two has a high salience value. Component four (Figure 3.13(c)) shows the same random variation reflecting noise in the data as component four of the previous PARAFAC model.



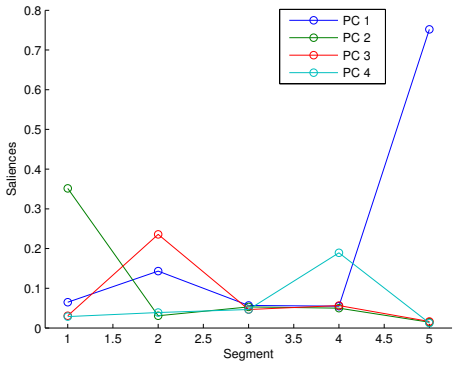
(a) Scores: q_1 vs. q_2



(b) Scores: q_1 vs. q_3



(c) Scores: q_1 vs. q_4



(d) Saliences: q_1 to q_4

Figure 3.13: Scores and saliences (weights of blocks/segments) of CCSWA on the three-way array of the segmented and mathematically transformed artificial GC-MS dataset with shifted peaks and noise. Samples are coloured according to Table 3.1.

3.5 Approach 2: 'SVD on each segment and PCA on eigenvalues'

The second approach was mainly inspired by the first one. The idea of segmenting chromatograms was kept, but the mathematical transformation of segments was changed. Each segment for each sample is decomposed using singular value decomposition (SVD) and only the first few singular values of each SVD are kept for further multivariate modelling using PCA.

3.5.1 Theoretical background

The basic idea behind SVD is the reduction of high dimensional data, such as large matrices with many variables, to a lower dimensional space which compromises the substructure of the data. SVD transforms correlated variables into fewer uncorrelated variables showing the various relationships among the original subjects (samples). In this way dimensions explaining most of the variation in the data are obtained. SVD can therefore also be understood as a data reduction method.

Many applications in signal processing and statistics make use of SVD. SVD is, for instance, the most often used algorithm for PCA. SVD consists of finding the eigenvalues and eigenvectors of the SSCP matrices XX^T and X^TX to obtain the left and right singular vectors (U an V), respectively, and the singular values in the diagonal of S, which are the square roots of the eigenvalues from XX^T or X^TX. Eigenvectors and eigenvalues exist in pairs meaning every eigenvector has a corresponding eigenvalue. Eigenvectors are new directions in the original data cloud, eigenvalues reflect the variance in the data in that direction. Singular values are becoming less important with descending indices. The first direction (component) explains therefore most of the variance in the data, the second direction the second most variance in the data and so on (Salkind, 2006). As SSCP matrices are used for the singular value

decomposition the same assumption on retention time shifts among samples hold as for approach 1 when SVD is applied to chromatographic segments of samples.

Approach 2 can be summarized as follows: After segmentation of the chromatograms (analogue to approach 1) each segment for each sample is decomposed using SVD, while only the first few singular values of each decomposition are used for further data analysis. The number of singular values to keep depends on the number of peaks in the segments (rank of the matrices). Note that for the sake of simplicity the segment size should be kept small similar to approach 1. The more similar segments are among samples the more similar are their decompositions. For instance replicates of samples show the same (or very similar) decomposition patterns, and have therefore the same (or very similar) singular values. For each sample all singular values of all samples are simply concatenated. In this way a matrix is obtained which after class centroid centering and scaling to intra-class variance can be analysed with PCA. A discussion on preprocessing of this matrix is presented in the next section.

The approach is summarized in matrix notation in the following. The segmentation of chromatograms is carried out according to equations 3.2 to 3.3 from approach 1. The SVD of X^{ik} is defined as follows:

$$X^{ik} = U^{ik} S^{ik} (V^{ik})^T, \tag{3.13}$$

where U^{ik} is an orthogonal $M \times M$ matrix, S^{ik} is a rectangular diagonal $M \times N_k$ matrix with non-negative entries and V^{ik} is an orthogonal $N_k \times N_k$-matrix. The M columns of U^{ik} and the N_k columns of V^{ik} are the left singular vectors and the right singular vectors of X^{ik}. The diagonal entries of S^{ik} are the so-called singular values $\sigma_1^{ik} \geq \cdots \geq \sigma_r^{ik} > 0$ of X^{ik}, where $r = \min\{M, N_k\}$.

Singular values are becoming less important with descending indices, we therefore only take the first Q singular values into consideration and represent them as a row

vector, that is

$$s^{ik} = (\sigma_1^{ik}, \ldots, \sigma_Q^{ik}), \tag{3.14}$$

for all $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Moreover, the row vectors s^{ik} are concatenated[3] to a row vector s_i over all K segments,

$$s_i = s^{i1} \frown s^{i2} \frown \ldots \frown s^{iK}, \tag{3.15}$$

for all $i = 1, \ldots, I$. Finally, we form the $(QK) \times I$-matrix Z by means of all row vectors $s_1, \ldots, s_I$ as follows:

$$Z = \begin{bmatrix} s_1 \\ \vdots \\ s_I \end{bmatrix}. \tag{3.16}$$

The final matrix Z is class centroid centred and scaled to intra-class variance before conducting PCA.

3.5.2 Application of approach 2 to the artificial GC-MS data set

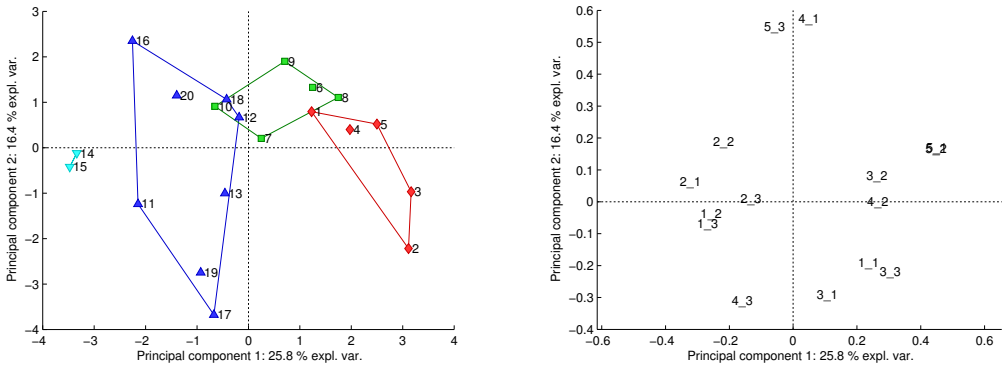
The performance of approach 2 on the artificial GC-MS data set without retention time shift and without baseline noise is shown in the following. The small random variation of peak sizes was however included to simulate a natural deviation of measurements. The number of singular values to keep were determined experimentally. Three singular values were kept for each segment. To show the impact of the preprocessing different PCAs with autoscaling and with class centroid centring and scaling by intra-class variance were conducted on the final matrix Z.

In standard mean centering the mean of each variable (column) in the data set is

[3]The concatenation $\frown$ of two arbitrary row vectors $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ is defined as:

$$x \frown y = (x_1, \ldots, x_n, y_1, \ldots, y_n).$$

calculated and removed (subtracted). For subsets (groups or classes) of samples class means can be calculated. The mean of these class means is the 'class centroid'. In class centroid centering this class centroids of each variable is calculated and removed (subtracted). Pooled variance (intra class variance) gives a weighted average of each group's variance. Scaling each variable by polled variance can be particularly interesting when the group variances are very unbalanced (larged differences). By this means class centroid centring and scaling by intra-class variance can be supportive at revealing differences among classes of samples.



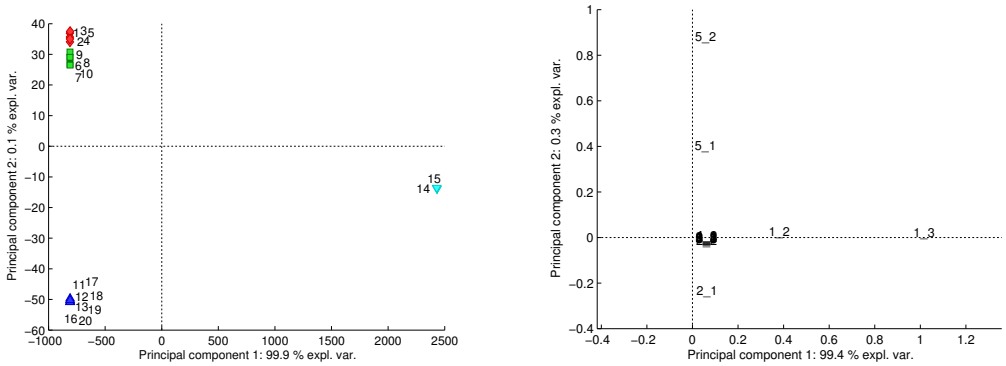
(a) Scores: PC 1 vs. PC 2

(b) Loadings: PC 1 vs. PC 2

Figure 3.14: Scores and loadings plots of the first two principal components of the PCA (autoscaled) on the final matrix Z of the artificial data set (without peak shifts). Samples are coloured according to Table 3.1. Numbers in the loadings plots refer to the segment and the singular value of the segment (e.g. 1_2: segment 1, second singular value).

Scores and loadings plots of PC1 (25.8 % explained variance) and PC2 (16.4 % explained variance) of the PCA on the autoscaled matrix are shown in figure 3.14. From the scores some structure among the samples can be observed, but no clear differentiation between all groups of samples is apparent. The same holds when only two singular values per segment are kept (data not shown). When class centroid centering and scaling to intra-class variance is applied for preprocessing of PCA clear separation between the groups of samples is obtained (Figure 3.15). Principal component 1 (99.9 % explained variance) reflects the differences of sample 14 and 15,

which solely contain peak 2 (segment 1). From the loadings plot (Figure 3.15(b)) it is evident that the second and third singular value of segment 1 are responsible for this difference. The difference of peak 9 in segment 5 between the first ten and the last ten samples is explained by PC2 (0.1 % explained variance). Accordingly, the fist two singular values of segment 5 show high loadings on PC2. Principal component 3 (Figure 3.16) explaining 0.1 % of variance reveals differences in segment 2 (peak 4) between samples.

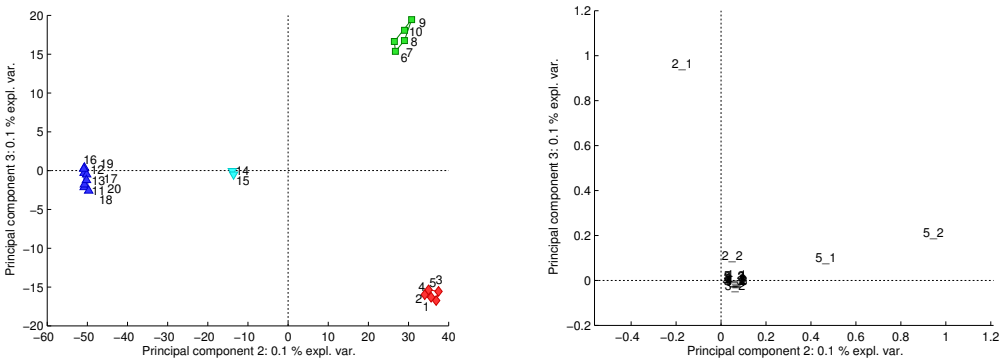


(a) Scores: PC 1 vs. PC 2 (b) Loadings: PC 1 vs. PC 2

Figure 3.15: Scores and loadings plots of the first two principal components of the PCA (class centroid centered and scaled to intra-class variance) on the final matrix Z of the artificial data set (without noise and peak shift). Samples are coloured according to Table 3.1. Numbers in the loadings plots refer to the segment and the singular value of the segment (e.g. 1_2: segment 1, second singular value).

The application of approach 2 on the artificial data set with and without noise and peak shifts resulted in similar groupings between samples. Results from PCA with class centroid centering and scaling to intra-class variance for the data set with noise and peak shifts are shown in Figure 3.17 and 3.18. In brief, PC1 reflects the differences of samples 14 and 15, PC2 shows the differences between the first ten and the last ten samples (Figure 3.17(a)). The difference of the samples one to five is explained by PC2 and PC3 (Figure 3.18(a)).

Approach 2 gives similar results to approach 1. However, approach 1 seems to be less sensitive to peak shifts than approach 2, as the variation inside the four

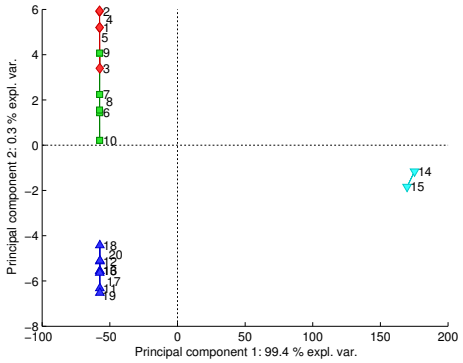


(a) Scores: PC 2 vs. PC 3

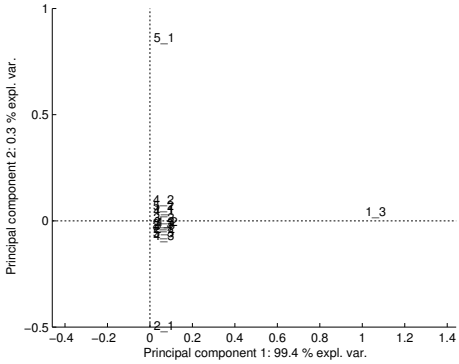
(b) Loadings: PC 2 vs. PC 3

Figure 3.16: Scores and loadings plots of principal components 2 and 3 of the PCA (class centroid centered and scaled to intra-class variance) on the final matrix Z of the artificial data set (without noise and peak shift). Samples are coloured according to Table 3.1. Numbers in the loadings plots refer to the segment and the singular value of the segment (e.g. 1_2: segment 1, second singular value).

groups of samples is smaller for approach 1 (Figures 3.12(a) and 3.12(a)) compared to approach 2 (Figures 3.17(a) and 3.18(a)). Moreover, approach 1 is unsupervised while for the class centroid centering and scaling to intra-class variance, the preprocessing of the PCA of approach 2 is supervised. Yet can approach 2 be seen as the 'simpler' approach, as the final PCA of approach 2 is easier, still provides interpretable results and is quicker to model compared to the PARAFAC model of approach 1.

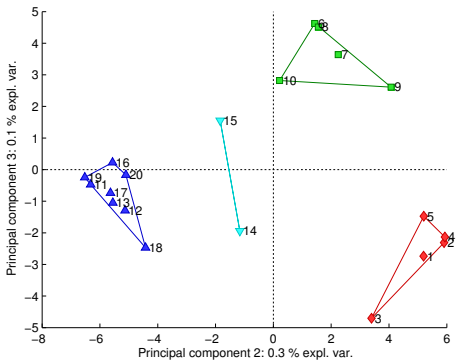


(a) Scores: PC 1 vs. PC 2

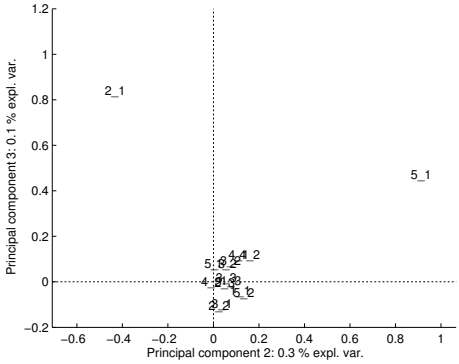


(b) Loadings: PC 1 vs. PC 2

Figure 3.17: Scores and loadings plots of the first two principal components of the PCA (class centroid centered and scaled to intra-class variance) on the final matrix Z of the artificial data set (with noise and peak shift). Samples are coloured according to Table 3.1. Numbers in the loadings plots refer to the segment and the singular value of the segment (e.g. 1_2: segment 1, second singular value).



(a) Scores: PC 2 vs. PC 3



(b) Loadings: PC 2 vs. PC 3

Figure 3.18: Scores and loadings plots of principal components 2 and 3 of the PCA (class centroid centered and scaled to intra-class variance) on the final matrix Z of the artificial data set (with noise and peak shift). Samples are coloured according to Table 3.1. Numbers in the loadings plots refer to the segment and the singular value of the segment (e.g. 1_2: segment 1, second singular value).

3.6 Application of the new data analysis approaches to experimental GC-MS data

The in Section 3.4 and 3.5 presented approaches are in the following tested on a set of real HS-SPME-GC-MS chromatograms of experimental wines.

3.6.1 Experimental

The data set explored in this study consists of solid phase microextraction (SPME) GC-MS analysis of Cabernet Sauvignon wines, which were fermented with different combinations of yeast and lactic acid bacteria using sequential inoculation and co-inoculation strategies.

3.6.1.1 Wine Samples

All wines were produced from the same Cabernet Sauvignon grapes from the 2012 vintage. Fermentations were carried out using six combinations of yeast and lactic acid bacteria commonly used in the wine industry to study their influence on the volatile composition of wines comparatively. Three wines were made with the yeast Lalvin Clos and the lactic acid bacteria Enoferm Alpha, Enoferm Beta and Lalvin PN4; two wines were made with the yeast Uvaferm RBS and the lactic acid bacteria Lalvin VP41 and O-Mega; and one wine was made with the yeast Uvaferm VRB and the lactic acid bacteria Enoferm Alpha (all from Lallemand Inc., Canada).

MLF is commonly conducted after alcoholic fermentation. However, alcoholic and malolactic fermentation can also be done simultaneously to save time and to prevent the risk of spoilage of the wine between the two fermentations. For this purpose, lactic acid bacteria are usually inoculated 24 h after yeast inoculation to conduct a simultaneous alcoholic and malolactic fermentation. This mode of inoculation is also called co-inoculation.

To obtain information on the differences of these two modes of inoculation all of the six yeast/bacteria combinations were fermented with sequential and co-inoculation of yeast and lactic acid bacteria. In total, the volatile composition of 12 experimental wines was studied here (Table 3.3). All yeast/bacteria combinations are commonly used in the wine industry. The major aim was to obtain analytical data of their impact on the volatile composition of wine.

Table 3.3: Cabernet Sauvignon wines. Sequential: lactic acid bacteria inoculation after completion of alcoholic fermentation; co-inoculation: lactic acid bacteria inoculation 24 h after yeast inoculation; LAB: lactic acid bacteria.

No.	Inoculation mode	Yeast starter culture	LAB starter culture	Abbreviation
1	co-inoculation	Lalvin Clos	Enoferm Alpha	clos alpha coin
2	sequential	Lalvin Clos	Enoferm Alpha	clos alpha seq
3	co-inoculation	Lalvin Clos	Enoferm Beta	clos beta coin
4	sequential	Lalvin Clos	Enoferm Beta	clos beta seq
5	co-inoculation	Lalvin Clos	Lalvin PN4	clos PN4 coin
6	sequential	Lalvin Clos	Lalvin PN4	clos PN4 seq
7	co-inoculation	Uvaferm RBS	Lalvin VP41	rbs VP41 coin
8	sequential	Uvaferm RBS	Lalvin VP41	rbs VP41 seq
9	co-inoculation	Uvaferm RBS	O-Mega	rbs 271 coin
10	sequential	Uvaferm RBS	O-Mega	rbs 271 seq
11	co-inoculation	Uvaferm VRB	Enoferm Alpha	vrb alpha coin
12	sequential	Uvaferm VRB	Enoferm Alpha	vrb alpha seq

3.6.1.2 HS-SPME-GC-MS Analysis

Headspace solid phase microextraction (HS-SPME) sampling was carried out in randomized order using a 100 µm polydimethylsiloxane (PDMS) fibre and the following procedure: 5 mL of the wine sample was transferred to a 20 mL headspace crimp-top vial and spiked with 152 µg L^{-1} ethyl hexanoate-d11 as internal standard. Two grams of sodium chloride (preheated to 250 °C and cooled to room temperature) were added and the vial was capped immediately using a PTFE-lined septum and aluminium cap. Each wine sample was submitted to HS-SPME sampling with agi-

tation at 500 rpm for 30 min. Fiber blank and column blank analyses were carried out regularly to confirm that no sample carry-over occurred. A standard 12 % hydro-alcoholic solution containing some esters and alcohols commonly present in wine was regularly analysed to monitor the performance of the system.

For GC-MS analysis an Agilent 6890 GC coupled to a quadrupole mass spectrometer Agilent 5973 N (Agilent Technologies, Palo Alto, CA, USA) was used applying electron impact ionisation (EI) at 70 eV. Full mass spectra were acquired in the range 35 u to 300 u at an acquisition rate of four spectra per second. The ion source temperature was set to 230 °C, and the detector voltage was 2105 V. Separation was carried out on a 30 m HP-5 MS column with an internal diameter (i.d.) of 0.25 mm and a film thickness of 0.25 µm. The following oven temperature program was used: 40 °C; kept for 5 min; ramped at 15 °C min^{-1} to 250 °C; and held for 5 min, resulting in a total run time of 25 min. Thermal desorption and injection were performed using a split/splitless injector, operated at 250 °C in the splitless mode, with a splitless time of 3 min. Helium was used as carrier gas at a constant flow of 1.0 mL min^{-1}. Linear retention indices were calculated using a series of *n*-alkanes. Experimental retention indices were compared to literature values to confirm tentative peak identification based on mass spectra. All chromatographic analyses were performed in triplicate.

3.6.1.3 Data Treatment

All raw chromatograms were exported from Agilent Chemstation version D.03.-00.611 as netCDF-files and imported into MATLAB version 8.0 (R2012b) (The MathWorks Inc., Natick, MA, USA) using built-in functions. All further data processing was done in MATLAB utilizing the freely available N-way toolbox (Andersson and Bro, 2000) and in-house written functions. Each of the 36 GC-MS raw chromatograms was transformed into a matrix of size 3977 × 266 (*elution profile* × *spectral profile*). Deconvoluted mass spectra were exported as ASCII text files in NIST .msp format

using an in-house written MATLAB function and imported into NIST 08 spectral library (Stein et al., 2008).

3.6.2 Application of approach 1 to experimental GC-MS data

The developed fingerprinting approach were applied to GC-MS data obtained for a set of twelve Carbernet Sauvignon wines fermented with different yeast/bacteria combinations using co-inoculation and sequential inoculation to study the impact of these factors on the volatile composition of the wines. SPME was chosen for sample preparation because of its simplicity for wine analysis in terms of full automation, speed and sensitivity (Vestner et al., 2011; Rocha et al., 2001; Antalick et al., 2010b). Although SPME fibres with mixed phases allow the extraction of a wider range of compounds, a PDMS fibre was chosen, as all PDMS degradation products contain silicone, which facilitates the differentiation of analytes from artefacts by means of siloxane fragments present in the mass spectra of the latter. This is particularly important when performing non-targeted analysis. A fast temperature ramp was used in this study to provide relatively fast GC separation. Under these conditions some resolution is sacrificed. However, the data analysis approach reported here takes the entire mass dimension into account, and therefore complete separation of peaks is not needed provided that co-eluting compounds differ in terms of their mass spectra. During the analyse of all samples, the system stability was monitored using a hydro-alcoholic standard solution containing common wine volatiles including ethyl butanoate until ethyl decanoate, butanol until decanol, isoamyl alcohol, isoamyl acetate, citronellol and nerolidol. Reproducibility of analyses were ensured using these monitoring injections. Matrix effects on the SPME extraction were not expected, as the composition of the analysed wines were very similar. Moreover, no significant changes of the absolute peak areas of the internal standard among samples have been observed (T-test, $\alpha = 0.05$ and n = 4 injections at beginning and end

of the sequence). The added internal standard was therefore not used to correct chromatograms. It should generally be noted that depending on the phase, analytes and matrix certain effects may occur during the SPME procedure such as analyte-matrix and analyte-sorbent interaction. For instance a direct consequence of coating saturation is inter-analyte competitive adsorption. SPME remains, however, a very powerful sample preparation technique for non-targeted analysis, but the SPME procedure has to be considered carefully regarding for instance matrix differences among samples (Souza-Silva et al., 2015; Gionfriddo et al., 2015).

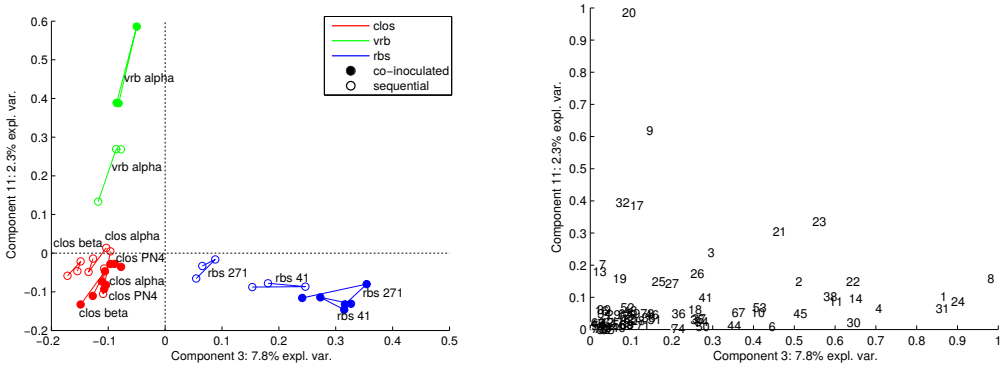
Initially, all chromatograms were divided into 84 small segments based on visual examination of overlays of total ion chromatograms (TICs) of all samples and of overlays of all mass channels for a single sample. Special attention was paid to avoid the inclusion of too many peaks in one segment and splitting of peaks into different segments. The latter is particularly important for segments containing peaks which shift between different samples. In this way, as few as possible peaks were included in each segment (one to five) and the dimensions of the segments ranged between 22 and 114 scans. The segments 15, 58 - 62, 72, 76, 77, 80, 81, 83 were excluded from the data set as they either contained only baseline or artefacts in the chromatograms. Seventy one small segments in total were kept for further analysis. To evaluate the effect of the number of segments, every two and every four neighbouring segments were combined which resulted in 36 and 18 bigger segments, respectively.

The outcome of the mathematical transformation (see section 3.4.1) of the segmented chromatographic raw data is a three-way array of size 36 $\times$ 36 $\times$ 71 (*samples* $\times$ *samples* $\times$ *number of segments*), 36 $\times$ 36 $\times$ 36 and 36 $\times$ 36 $\times$ 18, respectively. All arrays were mean centered across the first and second mode and scaled to unit variance within the third mode. The array which was obtained from the smallest segments (total of 71 segments) was analysed using CCSWA, Tucker3 and PARAFAC. CCSWA did not show any interpretable results against expectation (not shown). Tucker3 re-

sults were promising, although due to the nature of the Tucker3 model difficult to interpret. The results of the PARAFAC model were, however, much more informative and easier to interpret than the TUCKER3 results and revealed information on systematic differences among samples. The two other three-way arrays with 36 and 18 segments were therefore only analysed using PARAFAC. The number of components of the PARAFAC models were determined using the core consistency diagnostic (Bro and Kiers, 2003), by examination of residuals, and by evaluating captured variance and number of iterations until the PARAFAC algorithm converged for models with one to 20 components. For the three-way array with 71 segments a eleven component PARAFAC model was chosen, explaining 73.0 % of the total variation in the data set. The best PARAFAC models for the three-way array with 36 and 18 segments were a ten component PARAFAC model explaining 83.0 % of the total variation and a nine component PARAFAC model explaining 92.1 % of the total variation, respectively.

In general, PARAFAC loadings can be interpreted in the same way as PCA scores and loadings. In multi-way terminology, however, only the word 'loading' is used. For each mode of the analysed multi-way array a loading matrix is obtained. In the approach presented here, the first and second modes of the obtained PARAFAC model are identical, as the SSCP matrices from equation 3.11, which were compiled into a three-way array in equation 3.12, are symmetric. Congruence loadings were calculated for the third mode (segment mode) and each segment with a congruence loading value higher than 0.5 was considered as a 'moderate to strong correlated' with the raw data. Depending on the aim of the study, this value can also be chosen lower (e.g. 0.3, 'weak correlation') or higher (e.g. 0.7 'strong correlation'). A higher value for instance would be suitable if only highly correlated segments are of interest.

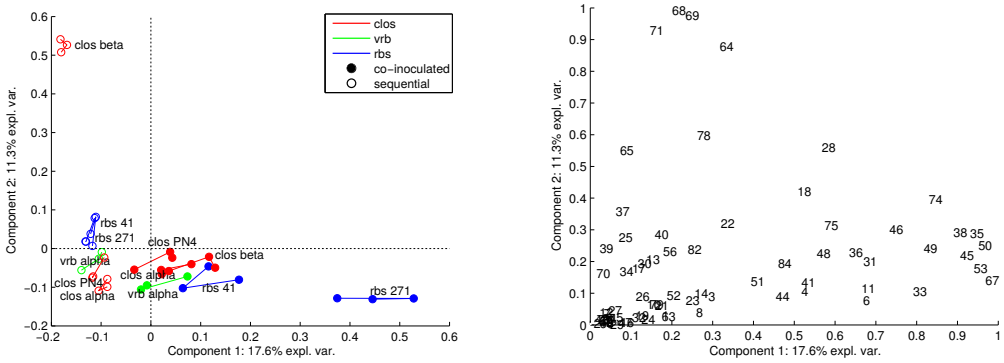
The information content of the three PARAFAC models are discussed and compared in the following. Examination of the loadings of the sample modes (first and second modes) of the PARAFAC model of the 71 segments showed that five of the



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

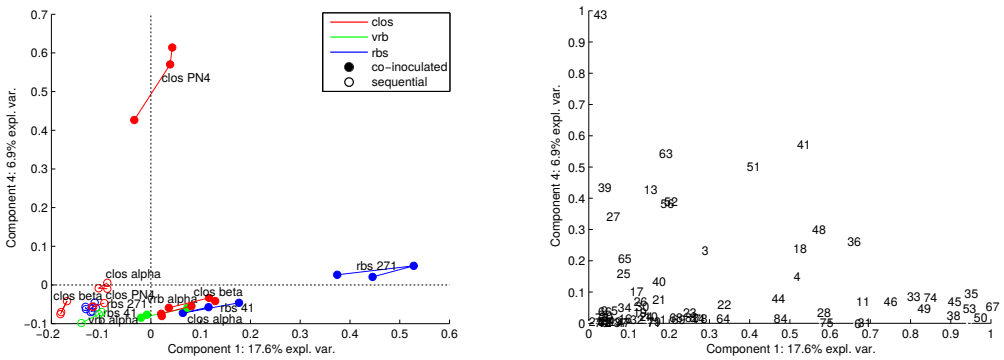
Figure 3.19: Loadings plots of PARAFAC components three vs. eleven (model with 71 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 3.20: Loadings plots of PARAFAC components one vs. three (model with 71 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



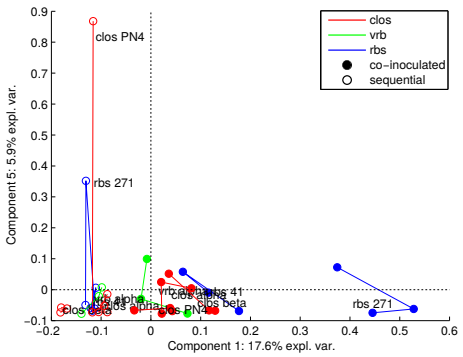
(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

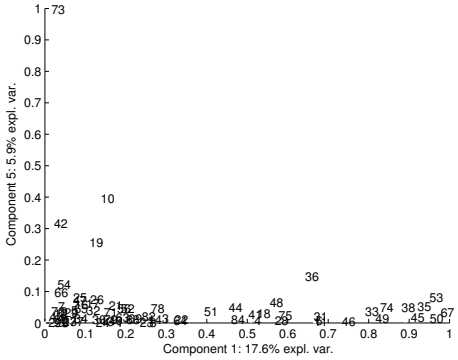
Figure 3.21: Loadings plots of PARAFAC components one vs. four (model with 71 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

eleven components contained important information revealing systematic differences between wines made with different yeast starter cultures and inoculation scenarios (Figures 3.19, 3.20 and 3.21). The remaining six components mainly reflect unsystematic variations in the chromatograms, for instance component five shown in Figure 3.22. From the congruence loadings of the segment mode of this component in Figure 3.22(b) it is evident that only one segment, that is segment 73, is responsible for the discrepancy of samples on this component (Figure 3.22(a)). The overlay of the TICs of segment 73 of all samples in Figure 3.22(c) shows that component 5 represents quantitative information in segment 73 very well. One injection of each of the wines made with the yeast/bacteria combination Lalvin Clos/Lalvin PN4 sequentially inoculated (clos PN4 seq) and the wine made with the yeast/bacteria combination Uvaferm RBS/O-Mega sequentially inoculated (rbs 271 seq) shows a much higher peak than all other samples in this segment. This pattern is exactly reproduced in the loadings of the sample mode of component 5. All other components containing redundant information are not further discussed here.

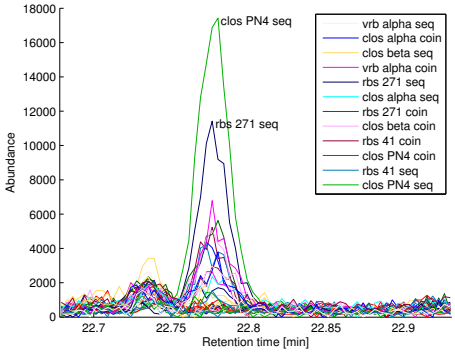
PARAFAC components three and eleven are displayed in Figure 3.19(a) showing



(a) First mode (samples) loadings



(b) Third mode (segments) congruence loadings



(c) TICs of segment 73

Figure 3.22: Loadings plots of PARAFAC components one vs. five (model with 71 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

the variation between wines fermented with different yeasts. Wines fermented with the yeast Uvaferm RBS (rbs) are separated from the wines fermented with the yeast Lalvin Clos (clos) and Uvaferm VRB (vrb) on component three (7.8 % explained variation), whereas the wines fermented with the yeast Uvaferm VRB differ from the other wines by component eleven (2.3 % explained variation). The impact of each segment on component three and eleven, respectively, is shown in the congruence loadings plots of the segment mode of these components in Figure 3.19(b). For component eleven only the segments 9 and 20 are responsible for the differences of the wines made with the yeast Uvaferm VRB compared to the wines made with the other two yeast starter cultures, considering congruence loading values higher than 0.5. The differences between the wines fermented with the yeast starter culture Uvaferm RBS and all other wines described by component three are correlated with the segments 1, 4, 8, 11, 14, 22, 23, 24, 30, 31 and 38.

Figure 3.20 shows the PARAFAC results for components one and two. Component one (17.6 % explained variation) mainly explains the differences in the wine fermented with the yeast Uvaferm RBS and the lactic acid bacteria O-Mega sequentially inoculated (rbs 271 seq), but this component also shows a general difference between co-inoculated and sequentially inoculated wines. Component two (11.3 % explained variation) mainly describes the distinction of the wine fermented with the yeast/bacteria combination Lalvin Clos/Enoferm Beta sequentially inoculated (clos beta) compared to all other wines. Congruence loadings of the segment mode for component one and two are shown in 3.20(b). Segments 4, 6, 11, 18, 28, 31, 33, 35, 36, 38, 41, 45, 46, 48, 49, 50, 53, 67, 74 and 75 had congruence loading higher than 0.5 on component one, while on component two segments 28, 64, 65, 68, 69, 71, 78 are important.

The relationship between the chromatographic raw data and the PARAFAC loadings of component two can be obtained by comparing the loadings in Figure 3.20

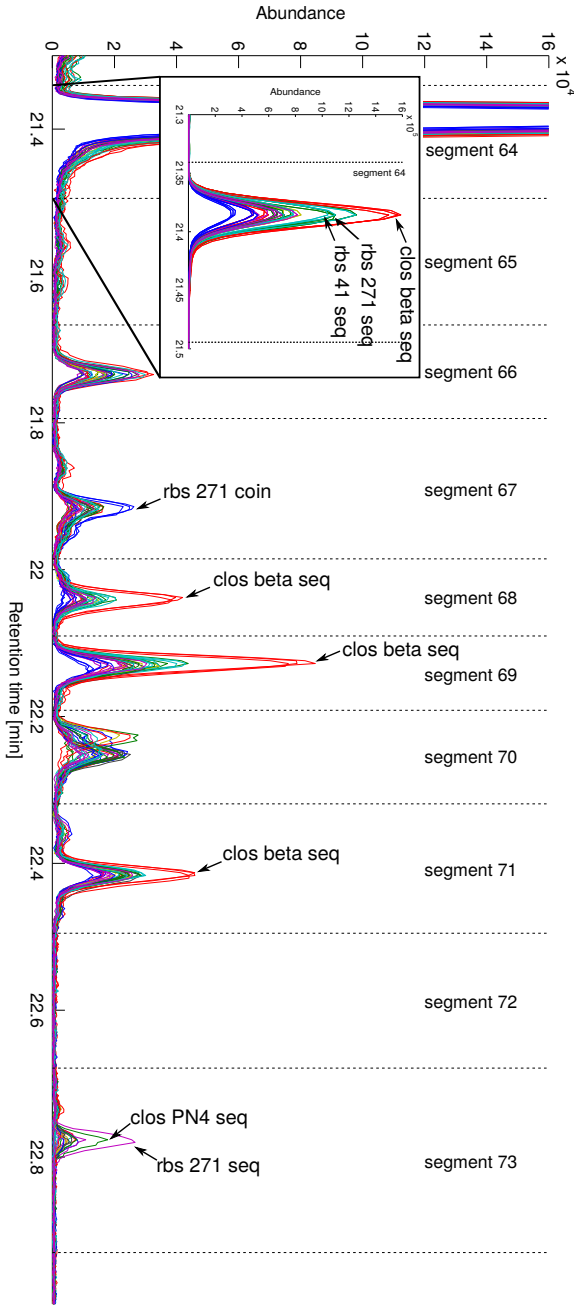
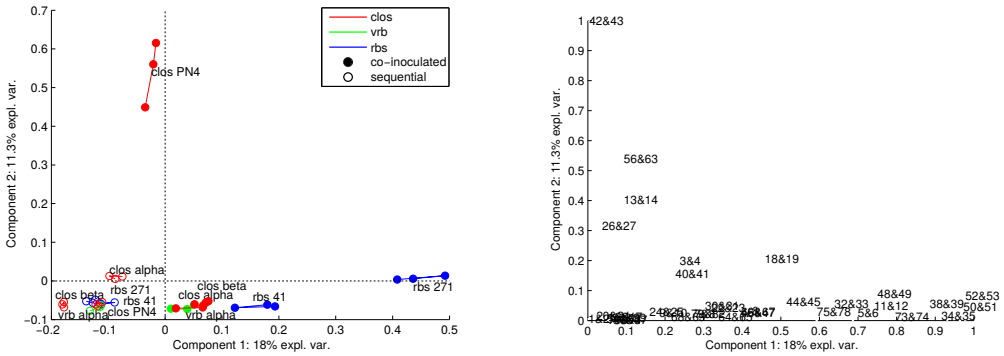


Figure 3.23: Overlays of total ion chromatograms (TICs) of all 36 injections (including replicates) of the HS-SPME-GC-MS analysis of the 12 Cabernet Sauvignon wines (segment 64 to segment 73).



(a) First mode (samples) loadings

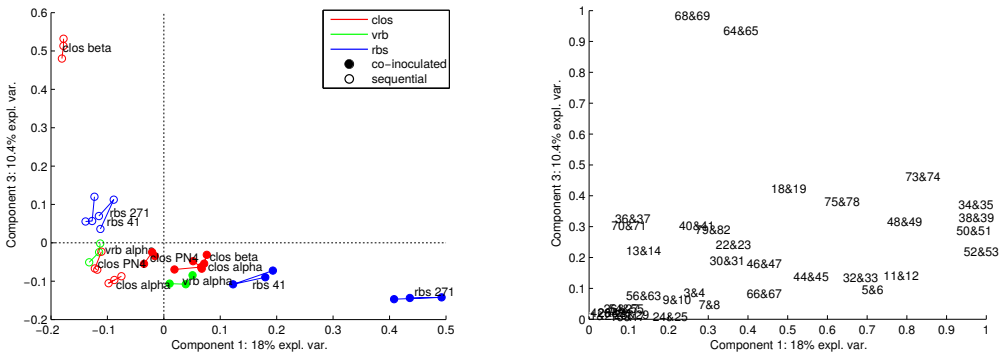
(b) Third mode (segments) congruence loadings

Figure 3.24: Loadings plots of PARAFAC components one vs. two (model with 36 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

with the TIC-overlays of all injections in Figure 3.23. The TIC overlays in Figure 3.23 confirm that the segments 64, 68, 69 and 71 (highest congruence loadings on PARAFAC component two; Figure 3.20) contain information on unique differences between the wine sequentially fermented with Lalvin Clos and Enoferm Beta (clos beta seq) and all other wines.

Component 4 explaining 6.9 % of the total variation in the data set differentiates the wine fermented with the yeast Lalvin Clos and the lactic acid bacteria Lalvin PN4 co-inoculated (clos PN4) from the other wines (Figure 3.21(a)). Responsible for this differences is especially segment 43, but also 41, 51 and 63 as shown in the congruence loading plot of the segment mode of this component (Figure 3.21(b)).

The results of the PARAFAC model with only 36 segments (neighbouring segments were combined) are very similar to the results of the PARAFAC model with 71 segments and will be discussed in the following. Component one of both PARAFAC models (Figure 3.20 and 3.24) reflect the same information, which is the differences of the wine fermented with the yeast Uvaferm RBS and the lactic acid bacteria O-Mega sequentially inoculated (rbs 271), and the difference between co-inoculated and



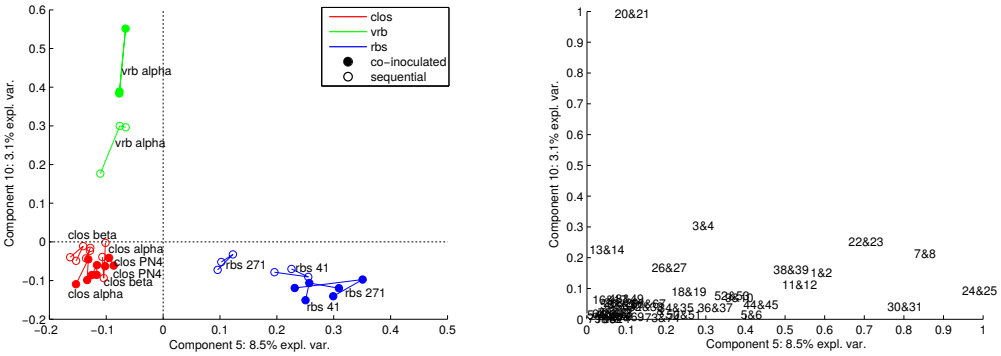
(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 3.25: Loadings plots of PARAFAC components one vs. three (model with 36 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

sequentially inoculated wines. Moreover, component three and two (Figure 3.24 and 3.25) of the PARAFAC model with 36 segments and component two and four (Figure 3.20 and 3.21) of the PARAFAC model with 71 segments show the same information on differences of the wines made with the yeast/lactic acid bacteria combination Lalvin Clos/Enoferm beta (clos beta) sequentially inoculated and Lavin Clos/Lalvin PN4 (clos PN4) co-inoculated, respectively. Components three and eleven of the PARAFAC model with the smallest segments (71 segments, Figure 3.19) reveal the same information as components five and ten (Figure 3.26) of the PARAFAC model with 36 segments, that is systematic differences according to the different yeast starter cultures.

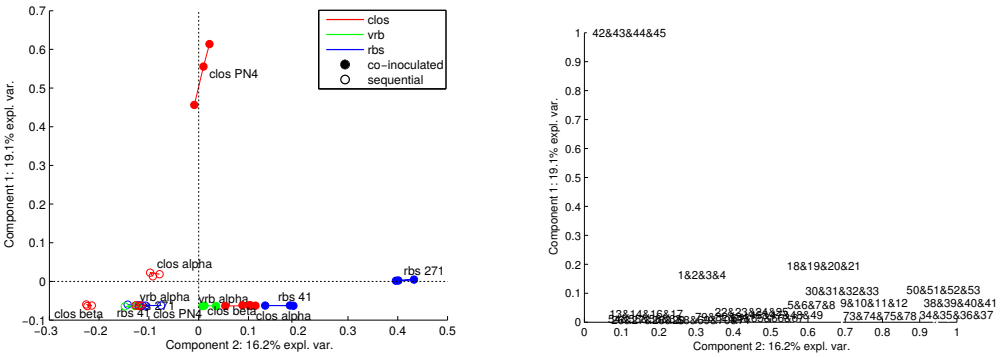
The results of the PARAFAC model where four neighbouring segments were combined (total of 18 segments) are not fully comparable to the results of the PARAFAC model with the smallest segments (71 segments). Only three components are comparable between these models. Component one (Figure 3.27) of the 18 segments PARAFAC model reflecting the differences between the wine fermented with the co-inoculated yeast Lalvin Clos and the lactic acid bacteria Lalvin PN4 (clos PN4) and



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 3.26: Loadings plots of PARAFAC components five vs. ten (model with 36 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



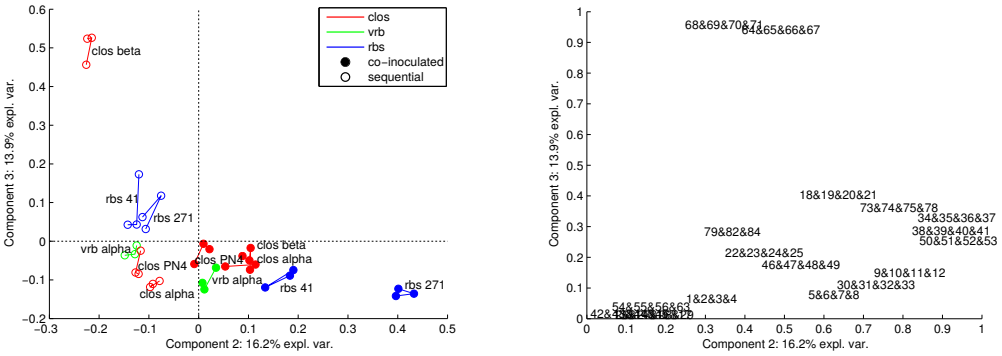
(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 3.27: Loadings plots of PARAFAC components two vs. one (model with 18 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

the other wines shows the same information as component 4 of the PARAFAC model with 71 segments. Component two (Figure 3.27) of the PARAFAC model with the biggest segments (18 segments) is comparable with component one of the 71 segment PARAFAC model mainly explaining the wine made with the yeast Uvaferm RBS and the lactic acid bacteria O-Mega (sequentially inoculated) and a tendency between co-inoculated and sequentially inoculated wines (Figure 3.20). Furthermore, component three of the PARAFAC model with 18 segments (Figure 3.28) shows differences of the wine made with sequential inoculation of the yeast Lalvin Clos and the lactic acid bacteria Enoferm Beta (clos beta) and is comparable with the information obtained from component two of the PARAFAC model with 71 segments (Figure 3.20). Information on the systematic differences caused by the yeast strains as obtained on component eleven and three (Figure 3.19) of the PARAFAC model with the smallest segments (71 segments) and on components ten and five (Figure 3.26) of the PARAFAC model with 36 segments could not be observed.

In conclusion, the comparison of the results of the three PARAFAC models with different segment sizes shows that the size of the segments clearly has an influence on the information obtained from the PARAFAC model. While the models with small and medium size (71 and 36 segments respectively) revealed the same information on systematic differences in the data, important information on systematic differences among the wines caused by the different yeast starter cultures could not be obtained from the PARAFAC model with the biggest segments (18 segments). These results demonstrate that smaller segments are beneficial. Another positive aspect of smaller segments is that they are easier to deconvolute afterwards.



(a) First mode (samples) loadings

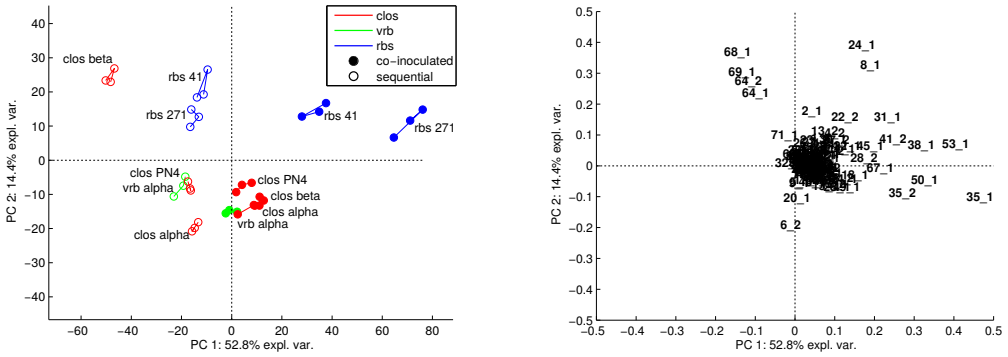
(b) Third mode (segments) congruence loadings

Figure 3.28: Loadings plots of PARAFAC components two vs. three (model with 18 segments); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

3.6.3 Application of approach 2 to experimental GC-MS data

To test approach 2 on the experimental data set the same segmentation as for the testing of approach 1 was used (see Section 3.6.2). Classes for class centroid centering and scaling to intra-class variance prior to PCA were first defined regarding the twelve treatments and subsequently regarding the three different yeast starter cultures used.

Initially, five different PCA models were tested for each scaling, where one to five singular values for each segment were kept. With the twelve treatments defined as groups the model with two singular values kept per segment revealed more information on the grouping of the samples compared to the first model (only one singular value per segment). The remainder of the models (more than 2 singular values kept per segment) did not reveal any extra information. For the models where classes were defined according to the three yeast starter cultures it was sufficient to keep only the first singular value of each segment.



(a) Scores: PC 1 vs. PC 2

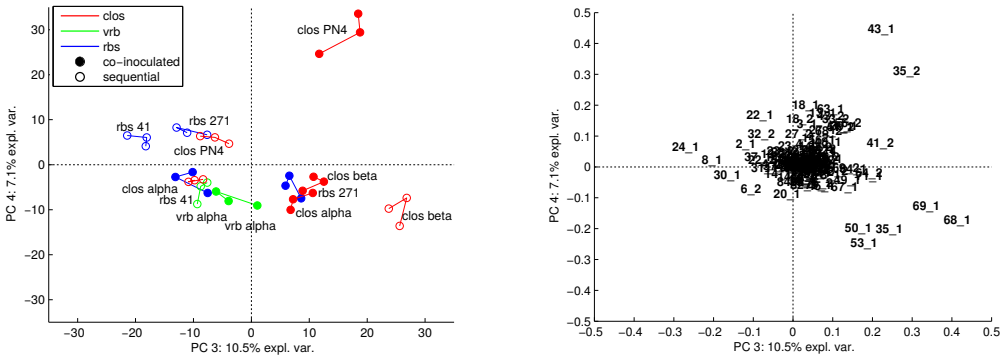
(b) Loadings: PC 1 vs. PC 2

Figure 3.29: Scores and loadings plots of PC1 and PC2 of the PCA on the final matrix Z (Equation 3.16) of the Cabernet Sauvignon data set, where each of the twelve treatments were used as classes for class centroid centering and scaling to intra-class variance. Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

The first six principal components of the PCA where the twelve different treatments were used as classes for the preprocessing contained important information on

differences between samples. PC1 explaining 52.8 % of variance (Figure 3.29) shows the difference between co-inoculated and sequentially inoculated wines. The segments 8, 24, 28, 31, 35, 38, 45, 50, 53 and 67 are mainly positively correlated with the co-inoculated wines. Similar information on the difference between co-inoculated and sequentially inoculated wines is revealed from PARAFAC component 1 of approach 1 (Figure 3.20). The PARAFAC model reveals however more segments contributing to this differentiation. The wines fermented with yeast Uvaferm RBS and the wine sequentially fermented with the yeast/bacteria combination of Lalvin Clos and Enoferm Beta (clos beta seq) are separated from all other wines on PC2 (14.4 % explained variance; Figure 3.29). The three sequentially inocluated wines with the yeast/bacteria combinations Uvaferm RBS/O-Mega (rbs 271), Uvaferm RBS/Lalvin VP41 (rbs 41) and Lalvin Clos/Enoferm Beta (clos beta) correlate with the segments 68, 69, 64, while the two wines co-inoculated with the yeast/bacteria combination Uvaferm RBS/O-Mega (rbs 271), Uvaferm RBS/Lalvin VP41 (rbs 41) correlate with the segments 24 and 8. The PARAFAC model of approach 1 reflects similar information on the difference of the sequentially inoculated yeast/bacteria combination Lalvin Clos/Enoferm Beta (clos beta) on component 2 (Figure 3.20) and on the differences of the wines fermented with the yeast Uvaferm RBS on component 5 (Figure 3.22). The information from the PARAFAC model of approach 1, however, shows this information on two separate components. Moreover, approach 1 gives more information of the importance of other segments for the observed groupings of samples, such as for segment 1, which contributes to the difference of the wines fermented with the yeast starter culture Uvaferm RBS.

PC3 and PC4 explaining 10.5 % and 7.1 % of variance, respectively, are displayed in Figure 3.30. The most interesting information on PC3 and PC4 are the correlation of the segments 35 and 43 with the wine co-inoculated with the Lalvin Clos and Lalvin PN4 (clos PN4) and the correlation of the segments 68 and 69 with the wine



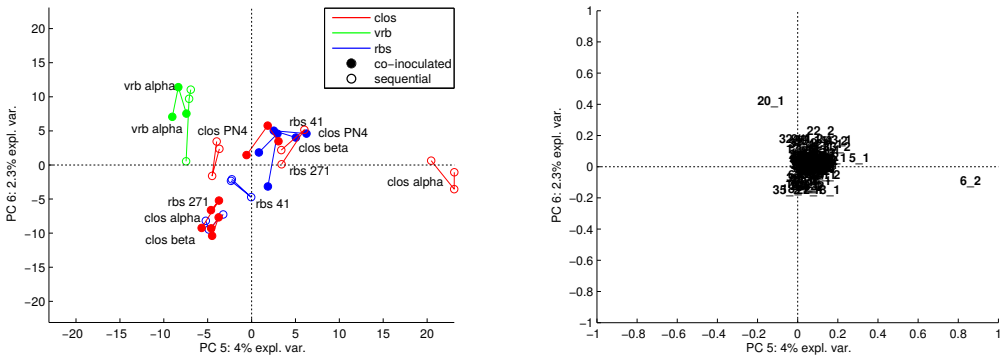
(a) Scores: PC 3 vs. PC 4

(b) Loadings: PC 3 vs. PC 4

Figure 3.30: Scores and loadings plots of PC3 and PC4 of the PCA on the final matrix Z (Equation 3.16) of the Cabernet Sauvignon data set, where each of the twelve treatments were used as classes for class centroid centering and scaling to intra-class variance. Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

sequentially inoculated with Lalvin Clos and Enoferm Beta (clos beta). The described information is similarly reflected in components 2 and 4 of the PARAFAC model of approach 1. However, PARAFAC component 2 also revealed the segments 64 and 71 to be important for the difference of the wine made with the yeast/bacteria combination Lalvin Clos/Enoferm Beta (sequential inoculation, clos beta seq) with the other wines. On the other hand, PARAFAC component 4 shows no contribution of segment 35 to the difference of the wine co-inoculated with the yeast/bacteria combination Lalvin Clos/Lalvin PN4 (clos PN4).

PC5 and PC6 (4.0 % and 2.3 % explained of variance) show the differences between the wines sequentially fermented with the yeast/bacteria combination Lalvin Clos/Enoferm Alpha (clos alpha) caused by segment 6 and the differences of the wines fermented with the yeast starter culture Uvaferm VRB caused by segment 20, respectively (Figure 3.31). The difference regarding the yeast starter culture Uvaferm VRB is also reflected in PARAFAC component 11 of approach 1. The difference of wines sequentially fermented with the yeast/bacteria combination Lalvin Clos/Enoferm Alpha is not explained by the PARAFAC model of approach 1.

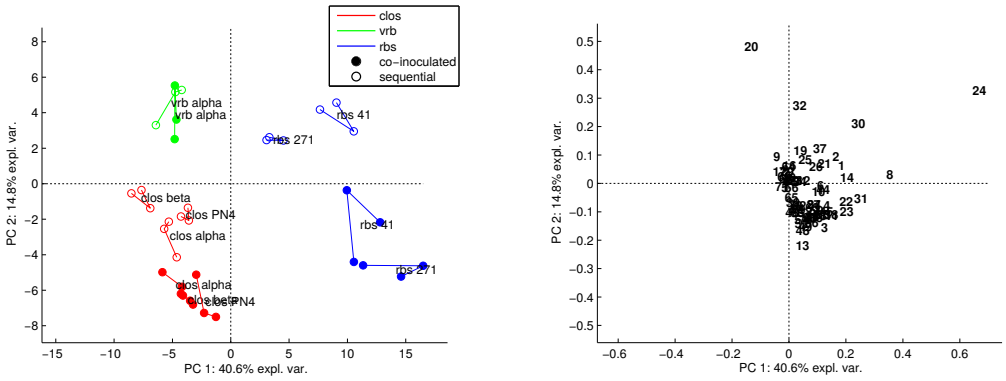


(a) Scores: PC 5 vs. PC 6

(b) Loadings: PC 5 vs. PC 6

Figure 3.31: Scores and loadings plots of PC5 and PC6 of the PCA on the final matrix Z (Equation 3.16) of the Cabernet Sauvignon data set, where each of the twelve treatments were used as classes for class centroid centering and scaling to intra-class variance. Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

The PCA where classes for the preprocessing were defined according to the three yeast starter cultures is discussed in the following. PC1 and PC2 explain 40.8 % and 14.8 % of variance in the data, respectively (Figure 3.32). PC1 separates the wines fermented with the yeast Uvaferm RBS from all other wines. PC2 separates the wines fermented with the yeasts Uvaferm VRB and Lalvin Clos. Moreover, PC2 shows a difference between co-inoculated and sequential inoculated wines fermented with the yeast Uvaferm RBS. The wines fermented with Uvaferm VRB correlate as expected with segment 20. Highest loadings on PC2 show the segments 8, 24, 30. Moreover the segments 1, 2, 14, 22, 23, 31 seem to contribute to the differences of the wines fermented with the yeast Uvaferm RBS. PARAFAC components 3 and 11 from approach 1 explain similar differences between the three starter cultures (Figure 3.19). The three important segments 8, 24 and 31 with highest congruence loadings on PARAFAC component 3 (Figure 3.19(b)) also have highest loadings on PC1 (Figure 3.32(b)). Evaluation of the importance of other segments from the loadings of PC2 is, however, difficult.



(a) Scores: PC 1 vs. PC 2

(b) Loadings: PC 1 vs. PC 2

Figure 3.32: Scores and loadings plots of PC1 and PC2 of the PCA on the final matrix Z (Equation 3.16) of the Cabernet Sauvignon data set, where classes for class centroid centering and scaling to intra-class variance were defined according to the three yeast starter cultures. Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

3.6.4 Approach 1 vs approach 2

The results of approach 1 and 2 are very similar. The PARAFAC results from approach 1 are however easier to interpret, especially in terms of the importance of segments for a certain grouping of samples. Moreover the groupings of samples observed from approach 1 appear to be clearer than the groupings from approach 2. PCA from approach 2, on the other side, is easier to model compared to the PARAFAC model regarding the evaluation of the correct number of PARAFAC components. Approach 2 revealed the difference of the wine sequentially fermented with the yeast-/bacteria combination Lalvin Clos/Enoferm Alpha (clos alpha seq), which was not detected using the unsupervised approach 1.

The fact that the difference of the wine sequentially fermented with the yeast-/bacteria combination Lalvin Clos/Enoferm Alpha (clos alpha seq) was not detected using approach 1 lead to the consideration of implementing class centroid centering and scaling to intra-class variance of each of the compilation matrices Y^k (Equation 3.9) into the algorithm of approach 1. The implementation of such a scaling step can bring out the differences between classes better, but it also makes approach 1 to a supervised method. The results of approach 1 with class centroid centered and to intra-class variance scaled compilation matrices Y^k (12 classes, one for each treatment) is shown in Figures B.1, B.2, B.3 and B.4 in appendix B.

3.6.5 Deconvolution and identification of compounds in important segments

In targeted analysis known and identified compounds are analysed. In non-targeted analysis, it is sometimes important to know the identity of compounds beforehand, but usually it is not known in advance. Only compounds which contribute to the differentiation of samples are identified (or tentatively identified) after statistical evaluation. For a more in-depth investigation of the data set all important

segments evaluated by approach 1 in Section 3.6.2 are more closely examined in the following.

From the discussion in Section 3.6.2 it can be summarized that the components one, two, three, four and eleven from the PARAFAC model with 71 segments are important to explain information on systematic differences between the wines. The segments with congruence loadings higher than 0.5, which can be considered as 'medium to high correlated', are segments 4, 6, 11, 18, 28, 31, 33, 35, 36, 38, 41, 45, 46, 48, 49, 50, 53, 67, 74 and 75 for component one, segments 28, 64, 65, 68, 69, 71 and 78 for component two, segments 1, 4, 8, 11, 14, 22, 23, 24, 30, 31 and 38 for component three, segment 41, 43, 51 and 63 for component four and segments 9 and 20 for component eleven. To confirm the results from PARAFAC modelling of the segmented and transformed GC-MS chromatograms and to study the important chromatogram segments which are responsible for the discrimination of samples in more detail, all of these 38 segments were deconvoluted using PARAFAC2 on each of the segments. The number of factors for each of the PARAFAC2 models were first evaluated as described by Johnsen et al. (2014) using the autochrom.m MATLAB function, which is kindly and freely provided on www.models.life.ku.dk (July 2014). The number of components of each model was then manually verified using the freely available N-way toolbox (Andersson and Bro, 2000) for MATLAB. The number of factors were checked, and if needed corrected, by examining core consistency, number of iterations until the algorithm converges, residuals, and the interpretability of the loadings. Moreover, non-negativity constraints were applied in the spectra mode. After exporting all deconvoluted mass spectra using an in-house written MATLAB function, tentative identification of the deconvoluted peaks were performed based on comparison of deconvoluted mass spectra with the NIST 08 spectral library. Furthermore, linear retention indices (LRI) were calculated using a homologous series of n-alkanes and compared with literature values to confirm tentative identifications.

Table 3.4: Summary of all segments showing high congruence loadings (> 0.5) on PARAFAC components one, two, three, four and eleven and details on the PARAFAC2 models of each segment with corresponding compounds.

	congruence loadings of PARAFAC component									
segment	1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
1			0.85			1	1	butanoic acid, 2-methyl-, ethyl ester (ethyl 2-methylbutyrate)	857	900
						2	2	butanoic acid, 3-methyl-, ethyl ester (ethyl 3-methylbutanoate)	861	852
						3	-	baseline		
4	0.51		0.69			1	7	acetic acid, hexyl ester (hexyl acetate)	1005	931
						2	8	propanoic acid, 3-methyl-, ethyl ester (iso-amyl iso-butyrate)	1003	812
						3	9	unknown m/z(%) = 69(100), 68(53), 142(32), 88(16), 97(12), 96(10)	999	
6	0.66					1	12	unknown m/z(%) = 57(100), 41(33), 43(27), 55(256), 70(242), 83(240), 56(215), 69(11)	1022	
						2	-	baseline		
						3	13	eucalyptol (1,8-cineole)	1025	877
8			0.97			1	15	2-hexenoic acid, ethyl ester (ethyl-2-hexenoate)	1048	860

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						2	-	baseline		
						3	-	baseline		
9					0.62	1	16	unknown m/z(%) = 125(100), 69(53), 41(24), 83(17), 95(17), 45(157), 55(146), 39(14)	1048	
						2	-	artefact (bleeding)		
						3	-	baseline		
						4	-	unknown m/z(%) = 71(100), 70(92), 43(66), 55(39), 41(28), 87(25), 89(23), 42(21)	1051	
11	0.67		0.59			1	19	propanoic acid 2-hydroxy-, 3-methylbutyl ester (isoamyl lactate)	1068	871
						2	20	1-octanol	1070	880
						3	21	unknown m/z(%) = 43(100), 55(69), 70(66), 41(63), 56(59), 69(44), 42(36), 84(31)	1069	
						4	22	acetophenone	1066	920
14			0.63			1	29	unknown m/z(%) = 70(100), 43(99), 57(85), 41(72), 85(70), 55(69), 71(58), 45(37)	1106	
						2	-	baseline		
						3	30	unknown m/z(%) = 131(100), 43(84), 132(62), 45(52), 55(50), 41(49), 44(46), 57(43)	1112	

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						4	31	unknown m/z(%) = 115(100), 43(56), 55(52), 45(38), 101(37), 85(31), 57(30), 41(29)	1111	
18	0.51					1	36	octanoic acid ethyl ester (ethyl octanoate)	1200	931
						2	-	baseline		
20					0.98	1	39	6-octen-1-ol, 3,7-dimethyl- (citronellol)	1231	888
						2	40	unknown m/z(%) = 41(100), 55(91), 69(87), 101(86), 43(73), 67(56), 45(55), 81(53)	1233	
						3	-	baseline		
22			0.63			1	42	hexanoic acid, 3-methylbutyl ester (isopentyl hexanoate)	1252	930
						2	43	hexanoic acid, 2-methylbutyl ester (2-methylbutyl hexanoate)	1255	868
						3	44	benzeneacetic acid, ethyl ester (ethyl benzeneacetate)	1250	852
						4	45	unknown m/z(%) = 70(100), 43(75), 71(60), 55(59), 99(55), 91(55), 41(46), 141(40)	1248	
						5	46	unknown m/z(%) = 121(100), 136(74), 93(62), 91(61), 70(51), 43(48), 41(27), 55(27)	1246	

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
23			0.54			1	47	acetic acid, 2-phenylethyl ester (phenylethyl acetate)	1262	961
						2	-	baseline		
						3	-	artefact (bleeding)		
						4	-	artefact (bleeding)		
						5	48	unknown m/z(%) = 117(100), 89(59), 94(48), 119(15)	1263	
24			0.89			1	49	unknown m/z(%) = 121(100), 136(71), 93(64), 91(20), 107(17), 43(16), 79(15), 77(148)	1268	
						2	-	artefact (bleeding)		
						3	50	nonanoic acid	1270	843
						4	-	baseline		
28	0.57	0.51				1	59	nonanoic acid, ethyl ester (ethyl nonanoate)	1297	892
						2	60	unknown m/z(%) = 96(100), 55(90), 41(90), 88(77), 138(71), 95(62), 67(50), 81(46)	1295	
						3	61	propyl octanoate	1294	841
30			0.63			1	65	unknown m/z(%) = 96(99), 55(90), 41(90), 88(77), 138(71), 95(62), 67(50), 81(46)	1331	

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						2	66	unknown m/z(%) = 117(100), 71(70), 43(40), 55(29), 89(26), 88(25), 101(24), 41(22)	1333	
						3	67	unknown m/z(%) = 99(100), 41(33), 69(30), 43(29), 71(27), 42(27), 101(22), 87(21)	1330	
						4	-	baseline		
31	0.67		0.85			1	68	octanoic acid, 2-methylpropyl ester (isobutyl octanoate)	1350	890
						2	69	unknown m/z(%) = 43(100), 57(77), 41(74), 55(70), 91(55), 44(51), 45(49), 56(36)	1352	
						3	-	baseline		
33	0.79					1	72	decanoic acid	1369	910
						2	-	baseline		
						3	73	unknown m/z(%) = 73(100), 60(84), 129(70), 55(69), 41(65), 43(62), 57(48), 71(46)	1368	
						4	74	naphthalene, 1,2-dihydro-1,1,6-trimethyl- (TDN)	1363	870
35	0.93					1	78	ethyl trans-4-decenoate	1389	873
						2	-	baseline		
						3	79	decanoic acid, ethyl ester (ethyl decanoate)	1397	

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
36	0.64					1	80	unknown m/z(%) = 163(100), 43(29), 91(19), 105(18), 121(179), 107(166), 93(15), 145(15)	1408	
						2	-	baseline		
						3	81	unknown m/z(%) = 43(100), 41(79), 55(78), 57(58), 69(56), 73(48), 44(46), 163(45)	1406	
						4	82	unknown m/z(%) = 73(100), 147(96), 43(26), 163(23), 41(20), 55(19), 45(16), 57(16)	1410	
						5	83	unknown m/z(%) = 43(100), 151(80), 109(79), 41(35), 55(24), 163(23), 69(22), 45(21)	1404	
						6	84	unknown m/z(%) = 69(100), 125(98), 43(83), 41(71), 73(68), 55(66), 163(60), 85(52)	1403	
38	0.89		0.57			1	87	octanoic acid, 3-methylbutyl ester (isoamyl octanoate)	1449	942
						2	88	unknown m/z(%) = 91(100), 127(46), 176(44), 103(31), 121(29), 92(22), 131(15), 77(11)	1450	
						3	89	octanoic acid, 2-methyl butyl ester	1451	921
41	0.52			0.57		1	-	baseline		
						2	96	unknown m/z(%) = 173(100), 155(67), 61(61), 175(47), 115(36), 60(31)	1490	

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						3	97	decanoic acid, propyl ester (propyl decanoate)	1492	857
						4	98	unknown m/z(%) = 55(100), 41(74), 155(39), 42(33), 133(22)	1493	
						5	99	unknown m/z(%) = 104(100), 57(76), 79(40), 143(39), 177(34), 53(31), 74(24)	1489	
43				0.99		1	102	unknown m/z(%) = 191(100), 192(14), 57(10), 41(4)	1515	
						2	103	butylated hydroxytoluene (BHT)	1520	959
						3	-	baseline		
						4	104	unknown m/z(%) = 192(100), 191(48), 43(47), 177(44), 91(41), 41(39), 73(37), 149(37)	1521	
						5	105	unknown m/z(%) = 145(100), 57(92), 177(64), 105(55), 91(48), 115(41), 41(41), 81(37)	1523	
45	0.91					1	109	unknown m/z(%) = 155(100), 57(90), 56(84), 173(67), 43(49), 182(47), 41(42), 55(33)	1547	
						2	110	unknown m/z(%) = 127(100), 155(74), 128(11), 181(10), 119(7), 156(7)	1550	
						3	-	baseline		

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
46	0.73					1	111	1,6,10-dodecatrien-3-ol, 3,7,11-trimethyl- (cis,trans-nerolidol)	1570	915
						2	112	unknown m/z(%) = 69(100), 55(99), 88(99), 41(94), 43(65), 60(47), 73(46), 67(46)	1571	
						3	-	baseline		
						4	-	artefact (bleeding)		
						5	113	unknown m/z(%) = 69(100), 41(90), 93(76), 43(73), 182(70), 55(66), 91(64), 71(53)	1574	
48	0.56					1	115	unknown m/z(%) = 43(100), 57(95), 145(85), 183(53), 55(51), 41(35)	1583	
						2	-	baseline		
						3	-	artefact (bleeding)		
49	0.82					1	116	unknown m/z(%) = 88(100), 55(84), 101(64), 97(51), 138(46), 96(45), 98(22), 110(21)	1588	
						2	-	baseline		
50	0.95					1	117	dodecanoic acid, ethyl ester (ethyl dodecanoate)	1595	971
						2	-	baseline		
51				0.5		1	-	baseline		

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						2	-	aretefact (bleeding)		
						3	118	unknown m/z(%) = 149(100), 177(27), 150(13), 105(10), 104(8), 176(8)	1610	
						4	119	unknown m/z(%) = 157(100), 167(54), 172(40), 132(35), 115(23), 158(19), 196(17)	1612	
53	0.94					1	121	pentadecanoic acid, 3-methylbutyl ester (iso-amyl decanoate)	1647	936
						2	122	unknown m/z(%) = 70(100), 155(71), 71(48), 173(46), 43(39), 104(38), 55(34), 41(22)	1650	
						3	-	baseline		
						4	-	artefact (bleeding)		
63				0.54		1	130	unknown m/z(%) = 55(100), 41(61), 69(58), 88(54), 101(43), 83(40), 97(39), 84(38)	1783	
						2	-	baseline		
64		0.66				1	131	tetradecanoic acid, ethyl ester (ethyl tetradecanoate)	1794	925
						2	-	baseline		
65		0.67				1	-	basline		
						2	-	artefact (bleeding)		

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						3	132	unknown m/z(%) = 120(100), 41(80), 55(75), 57(69), 138(65), 44(62), 121(60), 43(57)	1820	
						4	133	unknown m/z(%) = 73(100), 105(66), 55(32), 43(27), 44(26), 41(25), 147(25), 69(19)	1824	
67	0.96					1	135	dodecanoic acid, 3-methylbutyl ester (isoamyl laurate)	1847	891
						2	-	baseline		
						3	136	unknown m/z(%) = 70(100), 71(39), 43(33), 55(21), 183(17), 41(16), 69(14), 57(13)	1841	
68		0.51				1	137	unknown m/z(%) = 88(100), 101(63), 43(31), 55(24), 41(23), 157(18), 57(17), 73(17)	1859	
						2	138	unknown m/z(%) = 104(100), 105(27), 57(13), 44(11), 43(11), 41(11), 55(9)	1851	
						3	-	baseline		
69		0.57				1	139	unknown m/z(%) = 88(100), 101(66), 55(31), 41(28), 57(27), 43(25), 157(22), 69(20)	1866	
						2	-	artefact (bleeding)		
						3	-	baseline		
71		0.67				1	142	pentadecanoic acid, ethyl ester (ethyl pentadecanoate)	1896	874

Table 3.4 – continued

segment	congruence loadings of PARAFAC component 1	2	3	4	11	PARAFAC2 component no	no.	compound name	LRI[a]	MS match
						2	143	unknown m/z(%) = 100(100), 101(13), 55(12), 41(12), 44(12), 43(12), 88(11), 73(64)	1890	
						3	-	baseline		
74	0.83					1	146	ethyl 9-hexadecenoate	1976	917
						2	-	baseline		
75	0.57					1	147	hexadecanoic acid, ethyl ester (ethyl hexadecanoate)	1995	911
						2	-	baseline		
78		0.79				1	148	unknown m/z(%) = 88(100), 101(72), 55(32), 43(31), 41(31), 57(28), 69(19), 73(14)	2067	
						2	-	baseline		
						3	-	baseline		

[a]experimentally determined linear retention indices

Details on the PARAFAC2 models and the identified compounds are summarized in Table 3.4.

3.6.6 PCAs on deconvoluted peak areas

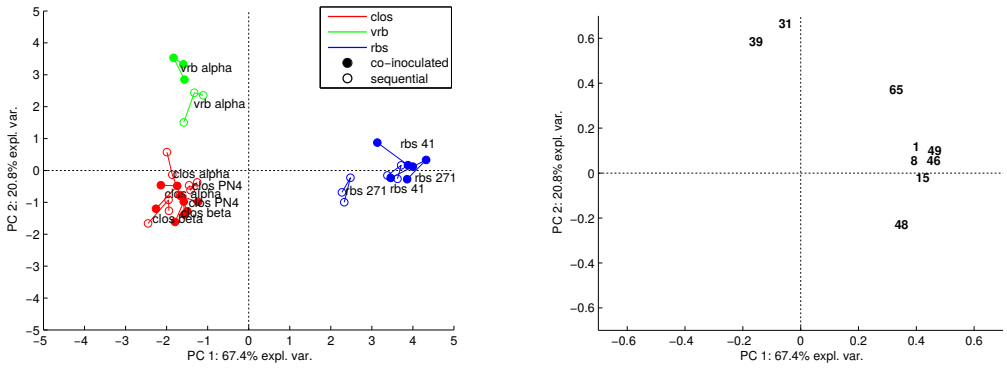
To visualize the above summarized and discussed results three different PCAs were constructed.

3.6.6.1 PCA 1: PARAFAC components 3 and 11

All compounds in the segments which had high congruence loadings on the components three and eleven of the PARAFAC model with 71 segments (Figure 3.19), which distinguished all samples according to the used yeast starter culture, were included in the first PCA. A two component PCA model was sufficient to separate the wines into three groups. The model was then improved by successively removing all compounds with low loadings on PC1 and PC2 (small impact on these two PCs). The wines fermented with the yeast starter culture Uvaferm RBS were separated from the other wines by PC1, explaining 67.4 % of the total variance (Figure 3.33(a)). The loadings in Figure 3.33(b) reveal that ethyl 2-methylbutyrate (1), isoamyl iso-butyrate (8), ethyl-2-hexenoate (15), the unknowns 46 and 49 (both terpenoid-like mass spectra) and the two unknowns 48 and 65 are positively correlated with the wines made with the yeast Uvaferm RBS. Moreover, the grouping of the wines fermented with yeast Uvaferm VRB is explained by PC2 (20.8 % explained variance). Citronellol (compound 39) and the unknown compound 31 are positively correlated on PC2 with these wines.

3.6.6.2 PCA 2: PARAFAC component 1

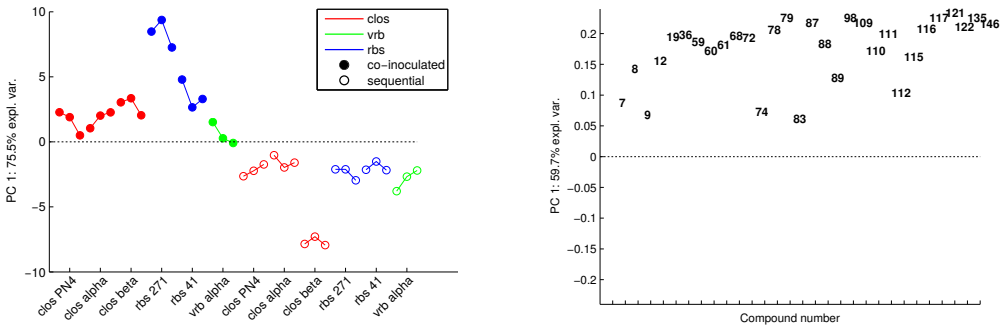
All compounds in the segments which had high congruence loadings on component one of the PARAFAC model with 71 segments (Figure 3.20) were included in



(a) Scores (b) Loadings

Figure 3.33: Scores and loadings plots of the PCA of compounds in segments which had high congruence loadings on components three and eleven of the PARAFAC modell with 71 segments; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

the second PCA. A one component model was sufficient to explain the differences between the co-inoculated wines and the sequentially inoculated wines. After successively removing all compounds with low loadings on PC1 (small impact on this component) a final one component model was obtained explaining 59.7 % of variance (Figure 3.34(a)). The branched esters isoamyl iso-butyrate (8), isoamyl lactate (19), isoamyl octanoate (87), isoamyl decanoate (121), isoamyl laurate (135) as well as isobutyl octanoate (68) and octanoic acid, 2-methylbutyl ester (89), the straight chain fatty acid ester ethyl octanoate (36), ethyl nonanoate (59), ethyl decanoate (79), ethyl deodecanoate (117), propyl octanoate (61), the two unsaturated ethyl trans-4-decenoate (78) and ethyl 9-hexadecenoate (146), the fatty acid decanoic acid (72), the terpenoid nerolidol (111), the unknown long chained fatty acid ester 122 and the unknowns 12, 60, 88, 109, 110, 115, 116 all correlate positively with the co-inoculated wines.



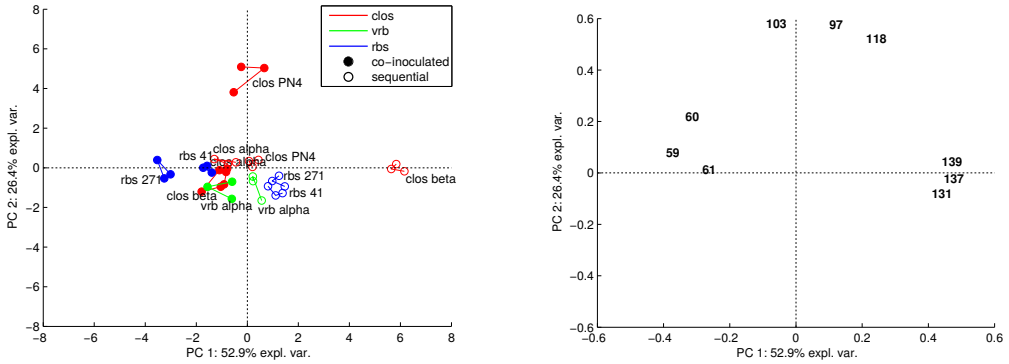
(a) Scores (b) Loadings

Figure 3.34: Scores and loadings plots of the PCA of compounds in segments which had high congruence loadings on component one of the PARAFAC model with 71 segments; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

3.6.6.3 PCA 3: PARAFAC components 2 and 4

The third PCA (Figure 3.35) included all compounds from segments which had high congruence loadings on the components two and four of the PARAFAC model with 71 segments (Figures 3.20 and 3.21). All compounds with low loadings (small impact on the model) were successively removed from the model. The wine made with the yeast Lalvin Clos and the lactic acid bacteria Enoferm Beta (sequentially inoculated, clos beta seq) is separated from all other wines on PC1, which explains 52.9 % variance. Ethyl tetradecanoate (131) and the two unknown long chain fatty acid ester 137 and 139 show positive correlation on PC1, while ethyl nonanoate (59), propyl octanoate (61) and unknown compound 60 correlate negatively with this PC. Principal component two (26.4 % explained variance) shows the difference of the wine which was co-inoculated with the yeast Lalvin Clos and the lactic acid bacteria Lalvin PN4 (clos PN4 coin). This difference is explained by propyl decanoate (97), BHT (103) and the unknown compound (118). BHT (103) and the unknown compound (118) are artefact compounds not associated to wine.

Several studies on the impact of the inoculation mode of malolactic fermentation



(a) Scores (b) Loadings

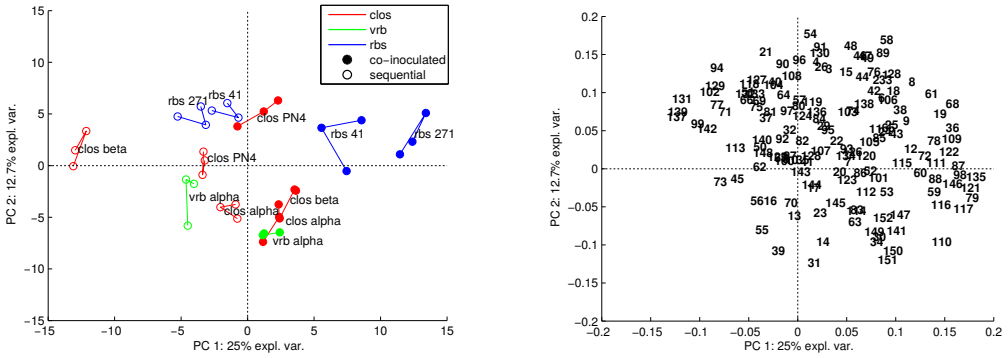
Figure 3.35: Scores and loadings plots of the PCA of compounds in segments which had high congruence loadings on components two and four of the PARAFAC model with 71 segments; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

and the yeast/lactic acid bacteria combination on the volatile composition of wine have been conducted, but no clear systematic changes have been reported (Antalick et al., 2010a; Gammacurta et al., 2014; Abrahamse and Bartowsky, 2012; Knoll et al., 2012). Some authors have observed higher amounts of some esters in co-inoculated wines (Abrahamse and Bartowsky, 2012; Knoll et al., 2012). Higher levels of long chain fatty acid esters as well as unsaturated and branched species as a function of malolactic fermentation inoculation mode have, however, not yet been reported. This is most likely due to the fact that long chain fatty acid esters are normally not the focus of targeted methods for general wine aroma analysis. Nevertheless, these compounds were included in the non-targeted approach used here, although this was *a priori* not specifically known.

3.7 Comparison of the new approaches to a reference method

As a reference method, PARAFAC2 was applied to all segments which have not been considered in the above discussed new approach and area values of all integrated deconvoluted peak profiles were analysed using PCA, according to Amigo et al. (2010a). A total of 152 peak area values were obtained in this manner. Figures 3.36 and 3.37 show the scores and loadings plots of PC1 (25.0 % explained variance), PC2 (12.7 % explained variance) and PC3 (11.8 % explained variance) of the autoscaled peak table. Note that only a relatively small proportion of variance is explained, even when compounds with low loadings were successfully removed (not shown). Some structural information is however revealed from the scores plots (Figures 3.36(a) and 3.37(a)), albeit the interpretation remains difficult.

PC1 shows, as component one from the PARAFAC model with 71 segments (Figure 3.20), a difference between most of the co-inoculated and sequentially inoculated wines. The co-inoculated wines fermented with the yeast starter culture Uvaferm RBS correlate most positively, while the wine made with the yeast starter culture Lalvin Clos sequentially inoculated with the Enoferm Beta (clos beta seq) correlates most negatively with this PC. The compounds 8, 12, 19, 36, 59, 60, 61, 68, 72, 78, 79, 87, 88 98 109, 101, 111, 115, 116, 117, 121, 122, 135 and 146 show high positive loadings on PC1 (Figure 3.36(b)). These results are comparable to component one of the PARAFAC model with 71 segments (Figure 3.34). While the compounds 131, 137 and 139 correlate negatively with PC1, showing a similar pattern as reflected in PARAFAC component two of the 71 segment model (Figure 3.35). PC2 shows differences of the wines fermented with the yeast starter culture Uvaferm RBS and the wine made with the yeast/lactic acid bacteria combination Lalvin Clos/Lalvin PN4 (co-inoculated, clos PN4 coin). This separation is however not very clear, while there is no valuable information extractable from the loadings plot (Figure 3.36(b)). Similar can be observed for PC3, which also explains differences of the wine made with the



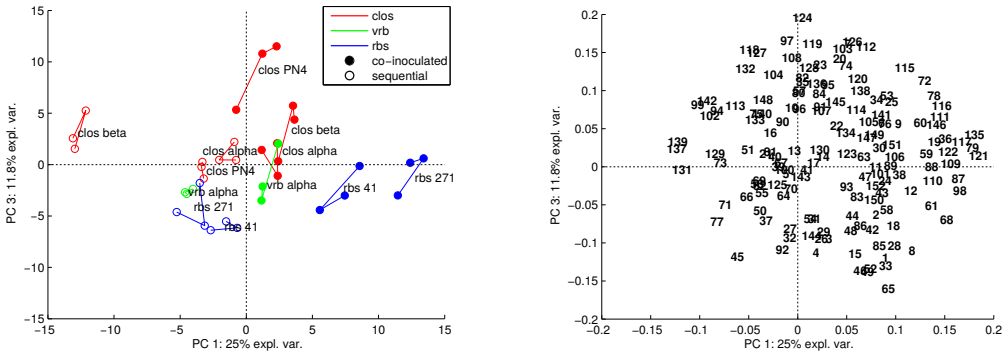
(a) Scores (b) Loadings

Figure 3.36: Scores and loadings plots of PC1 and PC2 of the PCA on all autoscaled compounds of all deconvoluted segments; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

yeast/lactic acid bacteria combination Lalvin Clos/Lalvin PN4 (co-inoculated, clos PN4 coin) and of the wines fermented with the yeast/lactic acid bacteria combination Uvaferm VRB/Enoferm alpha (co-inoculated, vrb alpha coin, Figure 3.37).

To obtain more information on the impact of the three yeast starter cultures a PCA on the whole peak table with class centroid centering and scaling to intra-class variance was constructed where classes were defined according to the three yeast starter cultures. Figure 3.38 shows the scores and loading of PC1 (38.8 % explained variance) and PC3 (10.7 % explained variance) of this PCA. The grouping according to yeast starter cultures are similar to the PCA on the autoscaled compounds of segments with high congruence loadings of component three and eleven of the PARAFAC model with 71 segments (Figure 3.33).

Overall the results from the multiple PCAs after the PARAFAC2 deconvolution of 38 important segments from approach 1 and the results from PCA after PARAFAC2 modelling of all 71 segments are comparable, albeit the latter were more difficult to interpret and more sophisticated methods then PCA with autoscaling are needed, such as supervised preprocessing (class centroid centering and scaling to intra class

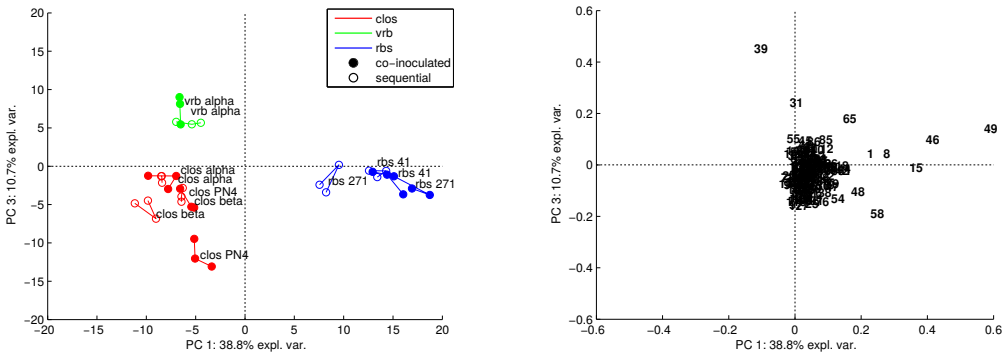


(a) Scores (b) Loadings

Figure 3.37: Scores and loadings plots of PC1 and PC3 of the PCA on all autoscaled compounds of all deconvoluted segments; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

variance).

The comparability of the results from the new approach using PARAFAC on segmented and mathematically transformed chromatograms in combination with PARAFAC2 deconvolution of important segments with subsequent PCA, and the deconvolution of all segments using PARAFAC2 and subsequent PCA modelling proves the validity of the results of the new approach. Only 38 segments of the chromatogram turned out to be important for the differentiation of samples using the new approach. Almost half of the 71 segments had to be deconvoluted using PARAFAC2, which is a considerable time saving. In this study only segments with congruence loadings greater than 0.5 were considered as 'medium to highly correlated' with the raw data. If, depending on the aim of a study, a higher value is chosen here, such as 0.75, which can be considered as 'highly correlated', even less PARAFAC2 models would have to be constructed and interpreted. The new approach can therefore be considered as a segment selection tool prior to (PARAFAC2) deconvolution of segmented chromatograms. Furthermore, the information on systematic differences obtained from the PARAFAC model on the segmented and transformed chromatograms can be used



(a) Scores (b) Loadings

Figure 3.38: Scores and loadings plots of PC1 and PC3 of the PCA on all compounds of all deconvoluted segments, where class centroid centering and scaling by intra-class variance was applied; Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

to study the important segments separately: separate PCAs can be constructed on only compounds from segments which are responsible for a certain grouping of samples. Peak tables obtained in this manner are much smaller than a global peak table from all compounds of the chromatograms and contain less redundant information. The PCAs constructed on these smaller peak tables are much easier to interpret, as has been shown above.

3.8 Conclusions

In this chapter, the two developed data processing approaches have been demonstrated as powerful techniques for the analysis of non-targeted GC-MS data. Both approaches were tested on artificial and real GC-MS chromatograms of multiple samples. The unsupervised approach 1 consists of three steps. First, all chromatograms are segmented and SSCP matrices are calculated for each segment and sample. This transformation of the chromatogram segments into SSCP matrices summarizes information on the variation and covariation of all mass channels in a segment and makes an alignment of peaks unnecessary. The following step, the compilation of the vectorized SSCP matrices into a compilation matrix for all samples in each segment and the transformation of these compilation matrices into SSCP matrices, gives information on the variation and covariation between samples in each segment as a function of the variation and covariation among mass channels in each segment. In the final step these SSCP matrices are merged to a three way array, which is then analysed using PARAFAC.

The supervised approach 2 also consists of three steps. Step one is, as for approach 1, the segmentation of the chromatogram. Step two is the singular value decomposition of every segment for every sample and the compilation of the first singular values of each segment into a final matrix. In the third step this matrix is class centroid centered and scaled to intra-class variance using predefined classes of samples and finally analysed using PCA.

A set of 36 chromatograms derived from triplicate HS-SPME-GC-MS analyses of twelve Carbernet Sauvignon wines was used to demonstrate the performance of the data treatment methodologies. Wines for instance could be differentiated according to yeast starter cultures and the inoculation mode of yeast and lactic acid bacteria. Approach 1 is more powerful in revealing clearer discrimination of samples by providing more structural information in the data compared to approach 2.

Compounds responsible for the discrimination of samples could be tentatively identified after deconvoluting peaks in the important segments using PARAFAC2. Based on the extra information obtained from the PARAFAC components of approach 1, multiple PCAs on the integrated deconvoluted signals of segments which are responsible for a certain grouping of samples provide in-depth insights to the observed phenomena.

The advantage of the novel GC-MS fingerprinting approach 1 presented herein could be confirmed by comparing it with PCA on peak areas from deconvoluted peak profiles of all chromatogram segments. A single PCA on the auto-scaled peak table of all deconvoluted compounds was however not sufficient to summarize all information obtained from the new approach, which underlines the advantage of the new approach 1. The new approache 1 is a fast alternative to conventional data analysis methods, as the only manual tasks are the segmentation of chromatograms and PARAFAC modelling. The new approach can also be seen as a segment pre-selection tool prior to deconvolution of chromatogram segments using e.g. PARAFAC2 or AMDIS.

CHAPTER IV

Application 1: Comparative aroma study on the impact of different malolactic fermentation scenarios on two Pinotage wine styles

4.1 Introduction

Aroma, which is characterized by volatile constituents, is one of the most important factors determining wine quality. The style of wine can be influenced by certain viticultural and oenological parameters, such as the harvesting date of grapes or different starter cultures for alcoholic and malolactic fermentation. Malolactic fermentation (MLF) is a crucial step particularly during the vinification of red wine. This second fermentation with lactic acid bacteria results in a natural deacidification, enhanced biological stability and improved mouth-feel of wine (Ribéreau-Gayon et al., 2006). The volatile composition, and as a consequence the sensory properties of wine, are also influenced during MLF. Besides buttery aroma caused by diacetyl, the aroma compound most associated with MLF, other aroma expressions have been reported to be influenced by MLF such as fruity, spicy, toasted and herbaceous notes. Clear trends how these notes develop as a function of MLF could however not yet been shown (Antalick et al., 2012; Gammacurta et al., 2014; Costello et al., 2012; Gámbaro et al., 2001; Sauvageot and Vivier, 1997; Mcdaniel et al., 1987).

Conventional descriptive profiling techniques such as Quantitative Descriptive Analysis (QDA) are usually performed for the sensory evaluation of experimental wines. These methods, however, require intensive training of panellists and are therefore time-consuming. Moreover, information about the importance of different attributes in the overall perception of panellists is not obtained. An alternative are rapid descriptive methods, which overcome this problem by letting the taster more freely decide how to indicate differences between samples. These fast methods, such as Projective Mapping, have recently gained more popularity. Napping can be seen as a special, restricted and defined case of Projective Mapping (Pagès, 2003; Dehlholm et al., 2012). Napping can be coupled with Ultra Flash Profiling to collect subjects semantic responses such as aroma descriptors, which can be collected as citation frequencies (Pagès, 2005a; Perrin et al., 2008).

In most aroma studies targeted approaches are applied, where a limited set of *a priori* known and identified compounds is accurately quantified. Considering, that samples can only be compared in terms of these selected compounds, targeted analysis can only confirm or reject an *a priori* assumption. On the other hand, non-targeted analysis are inherently more comprehensive by taking information of known and unknown compounds into account. By this means, a more holistic picture of the sample composition is obtained. Non-targeted approaches can consequently be more expedient in the search for compounds playing a key role in the differentiation of samples. Numerous agricultural and food related studies reflect an increasing interest in non-targeted analysis (De Vos et al., 2008; Cevallos-Cevallos et al., 2009; Croley et al., 2012; Wishart, 2008; Cubero-Leon et al., 2014).

The in Chapter III developed approach 1 is a fast and effective data analysis method for non-targeted GC-MS fingerprinting of wine volatiles. The goal of this chapter was the merging of fast GC-MS fingerprinting of wine volatiles with the rapid sensory screening method, partial projective mapping including a free choice

profiling of wines, to obtain an integrated picture of the sensory and chemical profile of experimental wines. Different MLF scenarios are compared to influence two different Pinotage styles: a fresh, fruity one made from early harvested grapes and a matured, full bodied one made from late harvested grapes. Commercial MLF starter cultures and the inoculation mode (co-inoculation or sequential inoculation) were chosen according common practices in commercial wineries. Two strategies for the merging of GC-MS and sensory data are evaluated: Quantitative and qualitative data matrices obtained from sensory evaluation and chemical fingerprinting are simultaneously analysed using multiple factor analysis (MFA) and the rotation of MFA scores from partial projective mapping onto the PARAFAC sample loadings using general procrusts analysis were evaluated.

4.2 Materials and methods

4.2.1 Wine making

To obtain two different Pinotage styles of the 2013 vintage grapes from the same vineyard were harvested at two different dates with differing sugar levels in the Stellenbosch region, South Africa. To obtain a modern, fruity Pinotage style and a full bodied Pinotage style grapes were harvested at 23.5 °B and 26.8 °B, respectively. After destemming and crushing, mashes of the early and late harvested grapes were aliquoted into three treatments with three replicates resulting in a 20 kg fermentation scale. Alcoholic fermentation was conducted using the yeast starter culture ICV-D80 (Lallemand Inc., Canada). $20\,\mathrm{g\,hL^{-1}}$ yeast starter culture were rehydrated with addition of $30\,\mathrm{g\,hL^{-1}}$ GoFerm Protect (Lallemand Inc., Canada) and inoculated according to the manufacturers instructions. Sequential inoculation, lactic acid bacteria inoculation after completion of alcoholic fermentation, and co-inoculation, lactic acid bacteria inoculation 24 h after yeast inoculation, were conducted using different com-

mercial MLF starter cultures. The following wines were made: co-inoculation with Lalvin VP41, co-inoculation with Lalvin V22 and sequential inoculation with Lalvin VP41 for the early harvested grapes and co-inoculation with Lalvin PN4, sequential inoculation with Lalvin PN4 and sequential inoculation with Lalvin VP41 for the late harvested grapes, respectively (see Table 4.1).

Table 4.1: Pinotage wines. Sequential: lactic acid bacteria inoculation after completion of alcoholic fermentation; co-inoculation: lactic acid bacteria inoculation 24 h after yeast inoculation; LAB: lactic acid bacteria.

No.	Harvesting time	Inoculation mode	LAB starter culture	Abbreviation
1	early	co-inoculation	Lalvin VP41	EH VP41 coin
2	early	co-inoculation	Lalvin V22	EH V22 coin
3	early	sequential	Lalvin VP41	EH VP41 seq
4	late	co-inoculation	Lalvin PN4	LH PN4 coin
5	late	sequential	Lalvin PN4	LH PN4 seq
6	late	sequential	Lalvin VP41	LH VP41 seq

4.2.1.1 HS-SPME-GC-MS Analysis

All GC-MS analyses were done 15 month after wine making. Headspace solid phase microextraction (HS-SPME) was carried out using a 100 µm polydimethylsiloxane (PDMS) fibre as follows: 5 mL wine sample (pH adjusted to 4.1 using sodium hydroxide solution) was transferred to a 20 mL headspace crimp-top vial and spiked with 152 µg L^{-1} ethyl hexanoate-d11 as internal standard. Two gram of sodium chloride (preheated to 250 °C) were added and the vial was capped immediately using a PTFE-lined septum and aluminium cap. HS-SPME sampling was done with agitation at 500 rpm for 30 min. Fiber blank and column blank analyses were carried out regularly after 8 injections to confirm that no sample carry-over occurred. To monitor the performance and stability of the system a standard 12 % hydro-alcoholic solution containing some esters and alcohols commonly present in wine (ethyl butanoate until ethyl decanoate, butanol until decanol, isoamyl alcohol, isoamyl acetate, citronellol

and nerolidol) was regularly analysed.

GC-MS analyses were carried out on an Agilent 6890 GC coupled to a quadrupole mass spectrometer Agilent 5973 N (Agilent Technologies, Palo Alto, CA, USA) using electron impact ionisation (EI) at 70 eV. Detector voltage and ion source temperature were set to 2105 V and 230 °C, respectively. Full mass spectra were acquired in the range from 35 u to 300 u at four spectra per second. For chromatographic separation a 30 m HP-5 MS column with an internal diameter (i.d.) of 0.25 mm and a film thickness of 0.25 µm was used. Thermal desorption and injection was done at 250 °C using a split/splitless injector in splitless mode, applying a splitless time of 3 min. The applied oven program was as follows: 40 °C; kept for 5 min; ramped at 15 °C min^{-1} to 250 °C; and held for 5 min. The total run time was 25 min. Helium was used as carrier gas at a constant flow of 1.0 mL min^{-1}. Linear retention indices were calculated using a series of *n*-alkanes. To confirm tentative peak identification based on mass spectra experimental retention indices were compared to literature values. All chromatographic analyses were performed in triplicate.

4.2.1.2 Data Treatment

GC-MS chromatograms were exported from Agilent Chemstation version D.03.-00.611 (Agilent Technologies, Palo Alto, CA, USA) as netCDF-files and imported into MATLAB version 8.0 (R2012b) (The MathWorks Inc., Natick, MA, USA) using built-in functions. MATLAB was used for all further data analysis. Moreover, the freely available N-way toolbox for MATLAB (Andersson and Bro, 2000) and in-house written MATLAB functions were used. Preprocessing of multi-way arrays was done using the nprocess.m function of the N-way toolbox (Andersson and Bro, 2000). Parts containing only baseline at the beginning and end of the chromatograms were removed. All GC-MS raw chromatograms were rearranged as matrices of size 3783 × 266 (*elution profile* × *spectral profile*). Deconvoluted mass spectra were ex-

ported as ASCII text files in NIST .msp format using an in-house written MATLAB function and imported into NIST 08 spectral library (Stein et al., 2008).

4.2.2 GC-MS fingerprinting: Segmentation, mathematical transformation and PARAFAC modelling of GC-MS chromatograms

The data analysis approach 1 for GC-MS fingerprinting described in Chapter 3.4 has been used here. A brief summary will be given in the following. By visually examining TIC overlays of all samples and overlays of all single ion chromatograms for some samples, GC-MS chromatograms are segmented along the retention axis into small sections containing a small number of peaks (approximately one to five peaks). Sums of squares and cross product (SSCP) matrices are calculated for every segment of each sample. Of each segment the upper triangular part of the obtained SSCP matrices are vectorized and concatenated into a compilation matrix. Subsequently, each of the compilation matrices are transformed into SSCP matrices, which are finally assembled into a three-way array. The final three-way array is of dimensions *number of samples* $\times$ *number of samples* $\times$ *number of segments* and can be decomposed using PARAFAC. The loadings of mode one and two (sample modes) are identical, as the SSCP matrices of the compilation matrices are symmetric. Systematic differences between samples can be determined by visual examination of the loadings of the sample mode (mode one and two). Congruence loadings (Lorho et al., 2006) of the segment mode can be used to identify the importance of segments responsible for the differences between samples. Subsequently, only segments, which contain information on interesting differences between samples, are further investigated. Congruence loadings with an value greater than 0.5 were considered as 'medium to high correlated'. Therefore, only segments with congruence loadings greater than 0.5 on selected PARAFAC components, which show systematic differences between samples, were investigated more in detail.

4.2.3 Deconvolution of important chromatogram segments and identification of compounds using AMDIS

A modification in this chapter to the described methodology in Chapter 3.4 is the deconvolution and identification of peaks in important chromatogram segments (congruence loadings greater than 0.5) with AMDIS (Stein, 1999). AMDIS has been used in numerous studies (Mallard, 2014; Fiehn, 2003; Koek et al., 2006; Halket et al., 1999; Meyer et al., 2010; Börner et al., 2007) for detection and deconvolution of GC-MS peaks prior to multivariate modelling. Although the PARAFAC2 approach of Section 3.6.5 of the previous chapter has been reported to be advantages in terms of greater resolution and sensitivity (Amigo et al., 2010a; Murphy et al., 2012), AMDIS was chosen here for the deconvolution as it is easier and faster to apply than the more time consuming PARAFAC2 approach. Moreover, the batch processing function of AMDIS enables automated processing of multiple chromatograms.

Deconvoluted mass spectra were compared with NIST08 library (Stein et al., 2008). Linear retention indices (LRI) were calculated using a homologous series of n-alkanes and compared with literature values to confirm tentative identifications. The batch processing function of AMDIS was used to integrate and export deconvoluted peak areas into text files (.txt). Peak tables obtained from AMDIS were further processed in MATLAB and R.

4.2.4 Partial projective mapping with free choice profiling

Sensory analysis was conducted in the same week of the GC-MS analysis. Partial projective mapping with free choice profiling (according to Ultra Flash Profiling as described by Perrin et al. (2008)) was performed with 18 wine experts from research laboratories of the Institut des Sciences de la Vigne et du Vin (ISVV), Bordeaux University. For orthonasal evaluation of the six experimental wines 50 mL of wine were presented in clear INAO wine glasses, which were labelled with random three-

digit codes and covered with plastic Petri dishes. The tasting was conducted at room temperature in an ISO 8589:2007 certified degustation room equipped with a cubicle for each taster. All six wines were simultaneously presented in random order to the assessors, which were asked to position wines which they perceive as similar close to each other and wines which they perceive as different apart from each other on a 42.0×59.4 cm sheet of paper. Moreover, all assessors were encouraged to write aroma descriptors of their own choice next to each wine.

For multivariate analysis the FactoMineR package (Lê et al., 2008) of the open source software R (version 3.1.1) was used. Multiple Factorial Analysis (MFA) was carried out with the x- and y-coordinates of the wines of each tasting sheet as a separate table (group) as has been described before (Pagès, 2005b). Aroma descriptors were counted for each wine and grouped together, in a way that for instance all red and black berry attributes were combined as 'red/black fruits'. All aroma descriptors which were named less than five times were excluded from further analysis. In this manner the following five groups were obtained: 'fruitiness', 'vegetal/herbaceous', 'red/black fruits', 'reductive' and 'lactic/butter'. The descriptor groups were included into MFA as a categorical supplementary table as has been descriped by Perrin et al. (2008).

4.3 Results and discussion

In this chapter results from the previously developed non-targeted GC-MS fingerprinting methodology (approach 1, Chapter III) of wine volatiles and results from fast projective mapping (including Ultra Flash Profiling) are integrated. The linkage of the information obtained from these two rapid methods facilitates a fast determination of correlations of volatile compounds with aroma descriptor groups. Different strategies of merging the results from Napping and the non-targeted GC-MS analysis will be discussed after the results of the chemical and sensory analysis are discussed

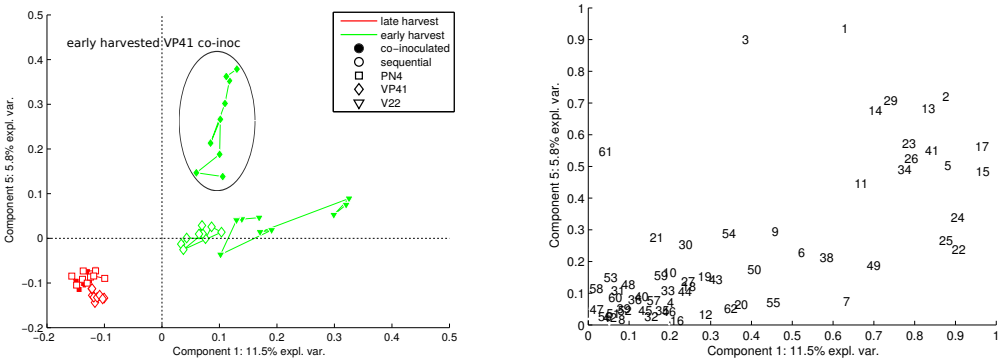
separately.

4.3.1 Fermentation performances

A fresh, fruity Pinotage style and a full-bodied Pinotage style were made from grapes of the same vineyard, which were picked at different harvesting dates with differing sugar contents. The wines from early harvested grapes completed alcoholic fermentation within 4 days. Malolactic fermentation in the co-inoculation fermentations with Lalvin VP41 finished 11 days after inoculation, while Lalvin V22 finished 21 days after inoculation. The Lalvin VP41 sequential inoculation finished MLF within 9 days. The starter culture V22 was inoculated at a dosage of $1\,\mathrm{g\,hL^{-1}}$, instead of $2\,\mathrm{g\,hL^{-1}}$ as in the manufactures instruction. As the viable cell numbers on day 3 were around $1 \times 10^5\,\mathrm{cfu\,mL^{-1}}$, a second inoculate of this treatment on day 7 with the $2\,\mathrm{g\,hL^{-1}}$ was performed. The initial cell numbers of the fermentations inoculated with VP41 were larger than $1 \times 10^6\,\mathrm{cfu\,mL^{-1}}$. The wines from late harvested grapes completed alcoholic fermentation within 6 days. The sequentially and co-inoculated wines with Lalvin PN4 completed MLF in 14 days after inoculation. The sequential inoculation with Lalvin VP41 finished MLF after 24 days. The initial cell numbers of the inoculated cultures in all the treatments were greater than $1 \times 10^6\,\mathrm{cfu\,mL^{-1}}$ did not lose any viability until the completion of MLF.

4.3.2 Non-targeted HS-SPME-GC-MS analysis

HS-SPME-GC-MS injections were performed in triplicate. Monitoring of the system stability was carried out throughout the full analysis time to assure the reproducible analyses of all samples, which is particularly important in non-targeted analysis. For this purpose blank injection and a hydro-alcoholic standard solution containing common wine volatiles were injected in regular intervals of eight samples. Chromatograms were normalized by the total peak area of the internal standard and

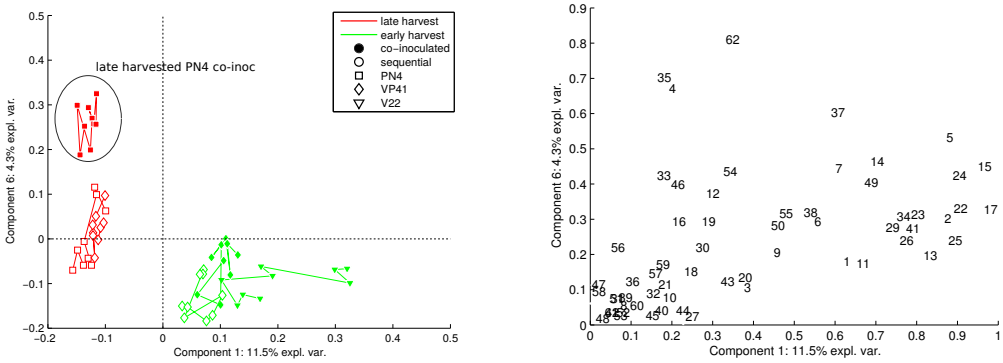


(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 4.1: PARAFAC loadings: component one vs. five. Numbers in (b) correspond to the segment number.

segmented by examining overlays of all total ion chromatograms (TIC) and overlays of all mass traces of some single injections. Special care was taken that not too many peaks were included into one segment and no peak was allowed to shift into a neighbouring segment. A total of 64 segments were defined in this manner. To examine the impact of the number of segments (segment size) neighbouring segments were combined in a second data set resulting in 32 segments. Mathematical transformation of the segmented chromatograms resulted in two three-way arrays with the dimensions $54 \times 54 \times 64$ and $54 \times 54 \times 32$. The first and the second mode of this array represent the wine samples including three technical replicates for each of the three biological replicates and mode three represents the chromatogram segments. To ascertain the correct number of components, multiple PARAFAC models with two to 20 components were calculated with ten repetitions to evaluate stability and convergence time of each model. Furthermore, core consistency diagnostic (Bro and Kiers, 2003), residuals, captured variance and convergence time of the algorithm were used to identify the correct number of components. The PARAFAC models of the $54 \times 54 \times 64$ array revealed more systematic differences compared to the models of the $54 \times 54 \times 32$ array. Any further discussion and representation of results is



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure 4.2: PARAFAC loadings: component one vs. six. Numbers in (b) correspond to the segment number.

therefore based on the PARAFAC model of the 54 × 54 × 64 array. All modes were checked for outliers using Hotelling's T-Square statistics and by examining residuals and loadings. Segments 28, 37, 63 and 64 were removed from the dataset. A 13 component PARAFAC model gave the best interpretable results by explaining 74.1 % of the total variation in the dataset.

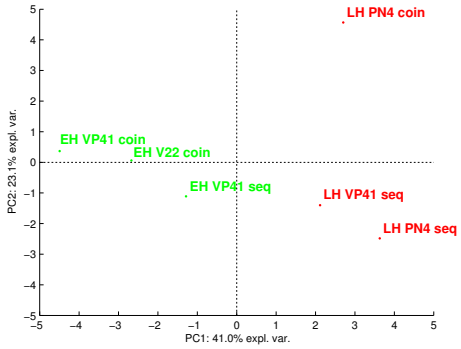
Examination of the loadings of the sample modes (first and second modes) of the PARAFAC model revealed that three components contain information on systematic differences, while other components reflect only non-systematic information. See Section 3.6.2 for a previous discussion on PARAFAC components which reflect only non-systematic differences among samples. All components explaining non-systematic structure in the data are not further discussed.

Component one explaining 11.5 % of variation reflects the difference between the wines made from early and late harvested grapes (Figure 4.1). The triplicate injections of one of the biological replicates of the wine from early harvested grapes co-inoculated with the MLF starter culture V22 differ from the other two biological replicates in terms of higher loadings on component one. These Differences are mainly explained by the segments 1, 2, 5, 6, 7, 11, 13, 14, 15, 17, 22, 23, 24, 25, 26, 29, 34,

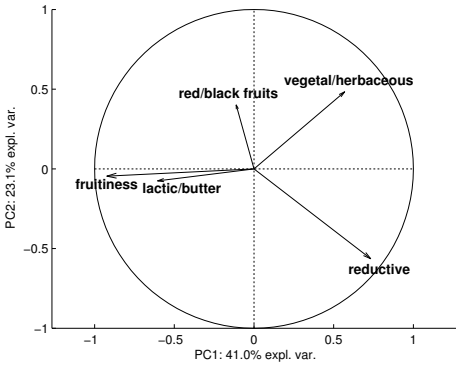
38 and 49 (congruence loadings larger than 0.5). Component five (5.8 % explained variation) separates the wine made from early harvested grapes co-inoculated with the MLF starter culture Lalvin VP41 from all other wines (Figure 4.1). The segments 1, 2, 3, 5, 13, 14, 23, 26, 29, 34, 37, 41 and 61 show high congruence loadings on component five. Component six (4.3 % explained variation) separates the wine made from late harvested grapes co-inoculated with MLF starter culture Lalvin PN4 from all other wines (Figure 4.2). The segments 4, 5, 35, 37 and 62 have congruence loadings larger than 0.5 on this component. To obtain more detailed information on compounds which are responsible for the differentiation of samples, all segments with congruence loadings larger then 0.5 were deconvoluted using AMDIS (Stein, 1999; Behrends et al., 2011). Significantly different (ANOVA, $\alpha = 0.05$, technical replicates were averaged) peak areas of deconvoluted peaks among the six wines were compiled in a peak table (see Table 4.2).

4.3.3 Sensory analysis

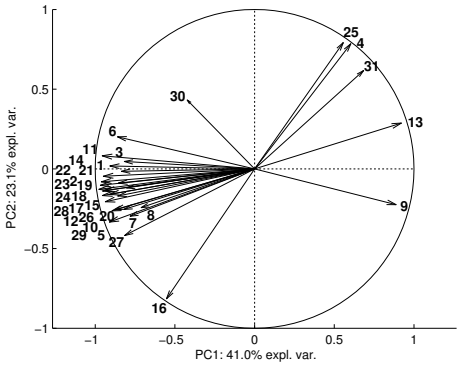
Partial projective mapping provides a holistic view on groupings and sensory characteristics of the tasted wines. As a member of the multi-block PCA family, MFA focuses on the analysis of several sets of variables (blocks or groups) which are collected on the same set of observations (samples). MFA is therefore the method of choice for the analysis of projective mapping data, especially when qualitative variables such as frequencies of sensory descriptor groups have to be incorporate into the analysis (Perrin et al., 2008; Pagès, 2005a). The representation of wines (scores of the global PCA) is shown in Figure 4.3(a). PC1 explaining 41.0 % of variance separates the wines according to early and late harvested wines. The second PC (23.1 % explained variance) shows differences between the wines from late harvested grapes co-inoculated with the MLF starter culture Lalvin PN4. The correlation of sensory descriptors is displayed in Figure 4.3(b). Overall fruitiness highly correlates with the



(a) Representation of wines (common factor scores)



(b) Representation of sensory descriptors (correlations)



(c) Representation of volatile compounds (correlations)

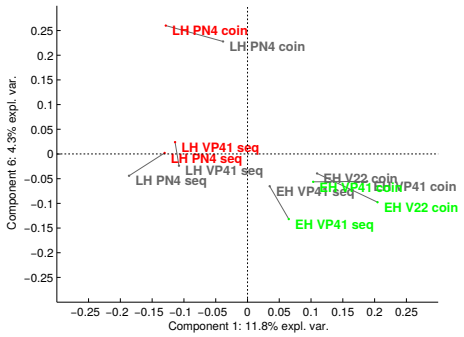
Figure 4.3: Results of MFA of partial projective mapping (orthonasal evaluation only), where frequencies of aroma descriptor groups of the free choice profiling were included as categorical supplementary variables (c) and peak areas (autoscaled) as continuous supplementary variables (b). Wines in (a) are labeled as follows: early harvested: EH (green), late harvested: LH (red), Lalvin PN4: PN4, Lalvin VP41: VP41, Lalvin V22: V22, co-inoculation: coin, sequential inoculation: seq. Numbers in (c) correspond to integrated compounds in Table 4.2.

early harvested wines with regards to the representation of the wine samples (Figure 4.3(a)). The descriptor reductive shows a strong correlation with the late harvested wines sequentially inoculated with the MLF starter cultures Lalvin PN4 and Lalvin VP41, respectively. The wine from late harvested grapes co-inoculated with Lalvin PN4 correlates with the descriptor vegetal/herbaceous. The descriptors lactic/butter and red/black fruits have only low correlation coefficients.

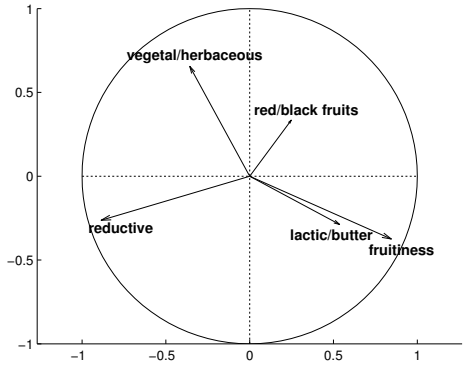
4.3.4 Merging of chemical and sensory data

To shed light onto the linkage of volatile compounds, assessors' ratings and sensory descriptors information of the fast sensory screening and the fast GC-MS screening of volatiles have to be merged. One possibility is the incorporation of the peak table obtained from GC-MS fingerprinting into the MFA of the partial projective mapping. The peak area values are, however, qualitatively different from the assessors' ratings and are therefore not supposed to be added as active elements into MFA, but can be projected as supplementary tables. Supplementary variables have no influence on the MFA, but they can be helpful for the interpretation of results and/or the linkage of other data. Descriptor frequencies are included as a supplementary table into MFA in the same way. The integration of the auto-scaled peak table into MFA of the partial projective mapping data is shown in Figure 4.3(c). A vast majority of compounds correlate negatively on PC1 with the wines obtained from early harvested grapes (Figure 4.3(a)) and the overall fruitiness (Figure 4.3(b)). These compounds are mainly branched fatty acid esters and acetates, which are known to contribute to fruity notes in wines (Ribéreau-Gayon et al., 2000). Unknown compound no. 25, unknown compound no. 4 with an terpenoid-like mass spectra and unknown compound no. 31 with the long chain fatty acid alike mass spectra are positively correlated, while the unknown compound no. 16 is negatively correlated with the wine made from late harvested grapes co-inoculated with the MLF-starter culture Lalvin PN4 and the aroma descriptor vegetal/herbaceous.

The direct linkage of the PARAFAC loadings with the results from MFA on the perceptual maps of all tasters and descriptor frequencies for each wine can also be of great importance, especially when certain sensory attributes are in the focus of a study. The main focus of this study for instance was the investigation of the impact of MLF on two Pinotage styles. MFA results clearly showed that the wine obtained from late harvested grapes co-inoculated with Lalvin PN4 solely correlates with veg-



(a) PARAFAC loadings of the sample mode (wines) with superimposed rotated MFA common factor scores (grey)



(b) Rotated representation of sensory descriptors (correlations)

Figure 4.4: (a) PARAFAC loadings of the sample mode (component 1 vs. 6) with superimposed rotated MFA scores (grey) of GPA. For MFA of the partial projective mapping the frequencies of aroma descriptors of the free choice profiling were included as categorical supplementary variables (c). Wines in (a) are labeled as follows: early harvested: EH (green), late harvested: LH (red), Lalvin PN4: PN4, Lalvin VP41: VP41, Lalvin V22: V22, co-inoculation: coin, sequential inoculation: seq. Numbers in (b) correspond to chromatogram segments in Table 4.2.

etal/herbaceous notes. GC-MS fingerprinting also revealed differences between this wine and the others. Merging of these results can be obtained by rotating the common factor scores of the MFA onto components of the PARAFAC loadings of the sample mode using general procrustes analysis (GPA). Note that the number of MFA components and PARAFAC components have to be the same. The representation of the sensory descriptors can subsequently be counter rotated according to the rotation of the common factor scores. This procedure was applied to the first and second principle component of the MFA results, which were rotated onto PARAFAC component one and six using GPA. A good match of the rotated common factor scores (PC1 and PC2) with the PARAFAC loadings of the segment mode (component 1 and 6) is shown in figure 4.4(a). The counter rotated representation of sensory descriptors from MFA is shown in Figure 4.4(b)). The obtained results are comparable with the above discussed MFA where the peak table of deconvoluted peaks was included into the MFA as supplementary table. The direct merging of PARAFAC results from

Table 4.2: Summary of all segments and their corresponding tentatively identified compounds showing high loadings (congruence loadings > 0.5) on PARAFAC components one, five and six.

	congruence loadings of PARAFAC component						
segment	1	5	6	no.	compound name[a]	LRI[b]	MS match
1	0.61	0.94		1	1-butanol, 3-methyl acetate (iso-amyl acetate)	815	969
2	0.86	0.73		2	3-hexenoic acid, ethyl ester	998	892
3		0.9		3	acetic acid, hexyl ester	1004	932
4			0.94	4	unknown m/z(%) = 93(100), 68(97), 67(88), 79(76), 94(56), 13(52), 92(51), 121(34)	1020	
5	0.87	0.51	0.73	5	2-hexenoic acid, ethyl ester	1039	839
6	0.54			6	unknown m/z(%) = 70(100), 87(97), 43(93), 71(88), 88(55), 102(47), 41(46), 55(36)	1053	
				7	ethyl 2-hydroxyhexanoate	1056	840
7	0.59			8	1-octanol	1070	909
11	0.66			9	unknown m/z(%) = 57(100), 81(52), 67(40), 56(39), 55(38), 41(30), 82(30), 83(28)	1219	
13	0.82	0.69		10	acetic acid, 2-phenylethyl ester (phenylethyl acetate)	1261	955
14	0.69	0.68		11	hexanoic acid, 3-methylbutyl ester (iso-amyl hexanoate)	1261	912

Table 4.2 – continued

segment	congruence loadings of PARAFAC component 1	5	6	no.	compound name[a]	LRI[b]	MS match
15	0.95			12	unknown m/z(%) = 121(100), 136(74), 93(61), 107(21), 91(20), 43(16), 73(15), 79(154)	1268	
				13	unknown m/z(%) = 125(100), 97(9), 126(99), 94(9)	1269	
17	0.97	0.57		14	unknown m/z(%) = 69(100), 41(81), 93(75), 192(62), 67(39), 121(37), 70(36), 99(35)	1289	
22	0.89			15	unknown m/z(%) = 101(100), 129(80), 57(19), 56(17), 102(10), 73(10), 41(8)	1332	
				16	unknown m/z(%) = 117(100), 71(83), 43(39), 88(24), 89(24), 83(19), 55(18), 57(11)	1334	
23	0.77	0.58		17	octanoic acid, 2-methylpropyl ester (iso-butyl octanoate)	1349	911
24	0.89			18	unknown m/z(%) = 159(100), 133(43), 119(42), 91(35), 220(29), 161(29), 73(28), 45(25)	1358	
				19	unknown m/z(%) = 43(100), 71(78), 73(64), 56(42), 85(37), 88(33), 41(32), 55(32)	1360	
25	0.88			20	unknown m/z(%) = 190(100), 107(62), 91(39), 105(38), 175(35), 93(29), 119(19), 77(18)	1429	
				21	unknown m/z(%) = 101(100), 129(82), 71(34), 70(30), 55(27), 43(25), 102(12), 41(8)	1431	

Table 4.2 – continued

	congruence loadings of PARAFAC component						
segment	1	5	6	no.	compound name[a]	LRI[b]	MS match
26	0.77	0.58		22	octanoic acid, 3-methylbutyl ester (iso-amyl octanoate)	1448	909
29	0.73	0.71		23	decanoic acid, propyl ester (propyl decanoate)	1492	831
34	0.75	0.58		24	decanoic acid, 2-methylpropyl ester (iso-butyl decaoate)	1547	848
35			0.9	25	unknown m/z(%) = 157 (100), 142(47), 141(23), 200(20), 156(16), 158(13), 115(8)	1558	921
37	0.59	0.51	0.51	26	unknown	1584	
38	0.53			27	dodecanoic acid ethyl ester (ethyl dodecanoate)	1596	895
41	0.78	0.6		28	pentadecanoic acid, 3-methylbutyl ester (iso-amyl pentadecanoate)	1646	924
49	0.67			29	unknown m/z(%) = 69(100), 70(53), 41(38), 81(33), 55(29), 136(24), 93(22), 43(21)	1847	
61		0.54		30	linoleic acid ethyl ester (ethyl linoleate)	2156	801
62			0.69	31	unknown m/z(%) = 88(100), 101(64), 43(26), 89(21), 55(20), 41(19), 157(19), 57(16)	2183	

[a]For each segment only compounds showing significantly different peak area values between treatments are listed.

[b]Experimentally determined linear retention indices.

GC-MS fingerprinting and MFA from partial projective mapping using GPA can be beneficial in terms of time when only selected informations are of a greater interest, for instance if in the here presented example the sole focus would have been on wines that correlate with vegetal/herbaceous notes, only the segments correlated with the wine made from late harvested grapes co-inoculated with Lalvin PN4 were needed to be deconvoluted and further investigated.

4.4 Conclusions

The new non-targeted data analysis approach (approach 1) was applied to study the impact of different MLF scenarios to a fresh, fruity and a full-bodied Pinotage style. Sensory evaluation of the wines was carried out using the rapid descriptive methods partial projective mapping with free choice profiling (Ultra Flash Profiling). By merging the results of the non-targeted GC-MS analysis of volatiles and the sensory data a more holistic overview of the aroma properties of the wines was obtained. Moreover, correlations of aroma descriptor groups and volatile compounds could be demonstrated. The attribute 'fruitiness' for instance showed a high correlation with the wines made from early harvested grapes, as well as with many ester, which are well known for their contribution to fruity notes in wine. Moreover, the wine made from late harvested grapes co-inoculated with the Lactic Acid Bacteria (LAB) starter culture Lalvin PN4 was rated as very different from the assessors than the other wines made from late harvested grapes. The descriptor 'vegetal/herbaceous' correlated highly with this wine, as well as with an unknown compound with an tepenoid-like mass spectra. From essential oils it is well known that terpenoid compounds can have 'herbaceous' aroma characteristics. The correlated compounds could therefore be responsible for the 'vegetal/herbaceous' note in the wine made from late harvested grapes co-inoculated with the LAB starter culture Lalvin PN4, but at least they are markers for the differences of this wine.

CHAPTER V

Application 2: Full factorial aroma study on the impact of grapevine age, yeast strain and must turbidity on the aroma of Riesling experimental wines

5.1 Introduction

It is widely assumed that the age of grapevines has a positive effect on the quality of the wine, but not much research has been conducted on this topic. In a recent study, on six red and white cultivars, wines from older grapevines generally had higher levels of titratable acidity (TA) and a better tannic structure compared to wines made from young grapevines (Zufferey and Maigre, 2008). Another study on Beihong wines showed an increase in the concentration of total volatiles and odour activity values (OAVs) for wines produced from older grapevines (Du et al., 2014). However, these differences might be more or less evident depending on the vintage (Reynolds et al., 2008). The impact of the age of grapevines on the sensory properties and composition of volatile compounds is still poorly understood.

Two very important oenological factors that are know to influence wine aroma are the yeast strain used for alcoholic fermentation and the degree of must turbidity prior

to fermentation. Must turbidity positively influences the fermentation rate and the final degree of fermentation. On the other hand, depending on the degree of turbidity the sensory characteristics and overall quality of wines can be negatively effected. Must clarification should therefore be effective, but not too drastic (Singleton et al., 1975; Groat and Ough, 1978; Losada et al., 2011; Williams et al., 1978; Ribéreau-Gayon et al., 2006). There is a broad agreement in literature that concentrations of higher alcohols increase as a function of must turbidity. The same applies for concentrations of ethyl and acetate esters (Houtman and Du Plessis, 1981; Nicolini et al., 2015; Losada et al., 2011; Karagiannis and Lanaridis, 2002). The degree of turbidity can also have an impact on varietal aroma compounds and glycoconjugates (Moio et al., 2004). Karagiannis and Lanaridis (2002) showed that the influence of turbidity on the volatile composition of wines also depends on the grape variety. The impact of the yeast strain on the sensory and volatile profile of wines has been intensively studied and reviewed (Rapp and Mandery, 1986; Romano et al., 2003; Swiegers et al., 2005; Lambrechts and Pretorius, 2000; Antonelli et al., 1999; Patel and Shibamoto, 2002). The conversion of must ingredients to sensorially important metabolites during fermentation such as acids, alcohols, carbonyl compounds, esters, sulfur compounds and monoterpenoids depends highly on the yeast strain (Swiegers et al., 2005).

The vast majority of studies focusing on the impact of oenological and viticultural practices on wine composition are conducted in a way that all parameters are kept constant and only the factor under study is varied. In this way interaction effects between factors are completely neglected, which can lead to biased conclusions. The usage of multifactorial design in experimental wine making is expedient, as it facilitates the evaluation of multiple factors at the same time. Full factorial designs test all possible conditions and can be used to find both main effects and interaction effects (Box et al., 2005).

In Chapter IV the multi-block PCA method MFA has been shown to be expedient to link data obtained from HS-SPME-GC-MS fingerprinting and sensory data from partial projective mapping with free choice profiling (Ultra Flash Profiling). Merging of data of these to fast methods provides an integrated view on and aroma of wines. The primary goal of this chapter was the combination of the strategy used in Chapter IV with a full factorial winemaking design for the detailed investigation of the impact of grapevine age, yeast strain and must turbidity and the dependencies among these factors (main and interaction effects) on the volatile composition and aroma of Riesling experimental wines. This combination will be shown to provide a powerful methodology for the detailed investigation of viticultural and enological factors influencing the aroma of wine.

5.2 Materials and methods

5.2.1 Viticulture

The experimental vineyard in Geisenheim, Germany, planted with *Vitis Vinifera* L. cv. Riesling vines of the clone 239-17 grafted on 5C Teleki rootstock. The grapevines were planted in 1971. In 1995 several rows were uprooted and replanted with grapevines of the same clone and rootstock. The result for the vintage 2013 is a vineyard with alternating blocks of vines that are 42 and 18 years old. Grapes of old and young grapevines were used separately for experimental wine making.

5.2.2 Experimental design and wine making

To obtain information on the effects of the three factors grapevine age, must turbidity and yeast strain on the volatile composition of Riesling wines, a 2×2×3 full factorial design was used for the experimental wine making. The structure of the full factorial design is presented in Table 5.1 and the 2×2×3 coded model matrix is

presented in Table 5.2. Multiple linear regression (MLR) was used to quantify main effects and one second order interaction effect.

The complete MLR model equation for the 3 factors (X_1, X_2 and X_3) is:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_{3A}X_{3A} + b_{3B}X_{3B} + b_{12}X_1X_2 \tag{5.1}$$

where Y is the response variable and b_0 is the constant term. The coefficients b_1, b_2 and b_3 account for the main effects of the factors X_1, X_2 and X_3, respectively. Moreover, the coefficient b_{12} represents the second order interaction term of the factors X_1 and X_2. Note that second order interaction effects of categorical factors with more than two levels (as the factor yeast X_3) cannot be calculated.

All twelve postulated fermentations in the model matrix (Table 5.2) were done in four replicates. Furthermore, duplicate HS-SPME-GC-MS analyses were performed for each of the fermentations, resulting in 96 analyses. All GC-MS injections were performed in random order.

Table 5.1: Structure of the 2×2×3 full factorial design used for the experimental wine making.

Factor	Level 1		Level 2		Level 3	
	uncoded	coded	uncoded	coded	uncoded	coded
X_1: age of grapevines	young	-1	old	+1		
X_2: must turbidity	clear	-1	turbid	+1		
X_3: yeast strain	Oenoferm KN	(1 0)	X5	(0 1)	EC1118	(0 0)

The grape must obtained from the two viticultural parcels (old grapevines 42 years; young grapevines 18 years) were pressed under the same conditions regarding press load and pressing program. Each of the two musts were divided into two lots for clarification and 30 mg L^{-1} SO_2 were added. Different Nephelometric Turbidity Unit (NTU) values of 300 and < 10 were adjusted with residual solids from must clarification after sedimentation for 24 h at 17 °C. Subsequently, 600 mL grape musts

were aliquoted into 750 mL bottles and inoculated with the commercially available *Saccharomyces cerevisiae* yeast strains Oenoferm Klosterneuburg (Erbslöh Geisenheim AG, Geisenheim, Germany) and Zymaflore X5 (Laffort, Bordeaux, France) and the *Saccharomyces bayanus* EC 1118 (Lallemand Inc., Canada) according to manufacturer's instructions. These yeast starter cultures were chosen because of different aromatic properties as stated in their product data sheets and from own experience. Bottles were closed with air locks before fermentation. All fermentation treatments were done in quadruplicate. Alcoholic fermentation took place at 17 °C and was monitored by determining loss of carbon dioxide. After fermentations were completed the wines were stored at 4 °C for one week and racked. Free SO_2 levels were adjusted to 50 mg L^{-1} during racking using potassium bisulfite. All wines were stored at 4 °C prior to chemical analysis. The common wine parameters glucose, fructose, fermentable sugars, density, tartaric acid, malic acid, total acidity, sugar free extract and glycerol were determined by FT-IR using a FOSS FT2 Winescan (FOSS Analytical A/S, Hillerød Denmark).

Table 5.2: Model matrix for the 2×2×3 full factorial design. Each of the 12 experiments were done in quadruplicate fermentations, resulting in 48 fermentations in total. Coding according to 5.1.

Experiments	b_0	b_1	b_2	b_{3A}	b_{3B}	b_{12}
1	1	-1	-1	0	0	1
2	1	-1	-1	0	1	1
3	1	-1	-1	1	0	1
4	1	-1	1	0	0	-1
5	1	-1	1	0	1	-1
6	1	-1	1	1	0	-1
7	1	1	-1	0	0	-1
8	1	1	-1	0	1	-1
9	1	1	-1	1	0	-1
10	1	1	1	0	0	1
11	1	1	1	0	1	1
12	1	1	1	1	0	1

5.2.3 HS-SPME-GC-MS analysis

All GC-MS analyses were conducted 17 month after winemaking. For headspace solid phase microextraction (HS-SPME) a 100 µm polydimethylsiloxane (PDMS) fibre was used. A standard SPME procedure typical for wine analysis was followed. Five millilitres of the wine sample were pipetted into a 20 mL headspace crimp-top vial together with two grams of sodium chloride (preheated to 250 °C and cooled to room temperature). The sample was spiked with 152 µg L^{-1} ethyl hexanoate-d11 as internal standard and the vial was capped immediately using a PTFE-lined septum and an aluminium cap. Each wine sample was extracted at 500 rpm for 10 min. To confirm that no sample carry-over occurred, fibre and column blanks were run regularly after eight injections. Moreover, a standard 12 % hydro-alcoholic solution containing some esters and alcohols commonly present in wine (including ethyl butanoate until ethyl decanoate, butanol until decanol, isoamyl alcohol, isoamyl acetate, citronellol and nerolidol) was regularly analysed to ensure constant and stable performance of the system.

GC-MS analyses were performed on an Agilent 6890 GC coupled to an Agilent 5970 N quadrupole mass spectrometer (Agilent Technologies, Palo Alto, CA, USA) operated in electron impact ionisation (EI) mode at 70 eV. A detector voltage of 2010 V was used and the ion source temperature was set to 230 °C. Full mass spectra were acquired in the range of 35 u to 350 u. For thermal desorption and injection a split/splitless injector operated at 250 °C with a splitless time of 3 min was used. Chromatographic separation was performed on a 30 m HP-5 MS column with an internal diameter (i.d.) of 0.25 mm and a film thickness of 0.25 µm. To guarantee a fast separation the GC oven temperature program was chosen as follows: 40 °C, kept for 5 min; ramped at 15 °C min^{-1} to 250 °C; and held for 5 min resulting in a total run time of 25 min. Helium was used as carrier gas at constant flow of 1.0 mL min^{-1}. Linear retention indices were determined using a series of *n*-alkanes and compared to

literature values to confirm tentative peak identification using the software AMDIS (Stein, 1999) based on deconvoluted mass spectra. Each GC-MS injection was performed in duplicate and all chromatographic analyses were run in random order.

5.2.4 GC-MS fingerprinting: Segmentation, mathematical transformation and PARAFAC modelling of GC-MS chromatograms

The data analysis approach 1 which has been described in Section 3.4 has been used here for GC-MS data analysis. The approach is summarized as follows: The initial step consists of the examination of overlays of total ion chromatograms (TICs) of all samples and the segmentation of the chromatograms (retention time profile $\times$ mass spectral dimension) along the retention time profile. Subsequently, all two-dimensional chromatogram segments of all samples are transformed to SSCP matrices in a way that the retention profile is eliminated. For each segment the upper triangular part of the obtained SSCP matrices are vectorized and concatenated into a compilation matrix. Subsequently, each of the compilation matrices are transformed into SSCP matrices, which are finally assembled into a three-way array of the size *number of samples* $\times$ *number of samples* $\times$ *number of segments*. The obtained three-way array is analysed using PARAFAC to find differences among samples and the corresponding chromatogram segments responsible for the discrimination of samples. PARAFAC can mathematically been seen as a three-way generalization of bilinear factor or component models such as PCA (Harshman and Lundy, 1994). The first and second modes of the obtained PARAFAC model represent the samples, similar to PCA scores. In multi-way terminology, however, only the word 'loading' is used. The modes one and two are identical, since the SSCP matrices which were compiled into a three-way array are symmetrical. The loadings of the third mode, representing the chromatogram segments, are provided as congruence loadings. Calculations were conducted using MATLAB version 8.0 (R2012b, The MathWorks Inc., Natick, MA,

USA) and the freely available N-way toolbox (Andersson and Bro, 2000).

5.2.5 Deconvolution of important chromatogram segments and identification of compounds using AMDIS

Peaks in chromatogram segments which had congruence loadings higher than 0.3 ('weak to strong correlation') were deconvoluted using the software AMDIS (Stein, 1999) for the same reason as described in Section 4.2.3. Deconvoluted mass spectra were compared with the NIST08 library (Stein et al., 2008). Linear retention indices (LRI) were calculated using a homologous series of n-alkanes and compared with literature values to confirm tentative identifications. Deconvoluted peak areas were obtained using the batch processing function of AMDIS and exported as .txt files for further data analysis.

5.2.6 Partial projective mapping with free choice profiling and multiple factor analysis (MFA)

Partial projective mapping with free choice profiling was performed in the same week as the GC-MS analyses with 18 wine experts from different research departments of the Hochschule Geisenheim University, Germany. Thirteen wines were examined; the twelve treatments plus an additional sample of the wine from turbid must of young grapevines fermented with Zymaflore X5 (X5rep_Y_T), which was used as a control sample to monitor the quality of obtained sample groupings. Fifty mL of every wine were presented in DIN Sensus wine tasting glasses (Zwiesel Kristallglas AG, Zwiesel, Germany). The glasses were labelled with random three-digit codes and covered with plastic Petri dishes for orthonasal evaluation. All wines were presented simultaneously in random order to the assessors. The tasting was conducted at room temperature in an ISO 8589:2007 certified tasting room equipped with a cubicle for each taster. The assessors were encouraged to position wines that they perceived as

similar close to each other and wines that they perceived as different away from each other on a 59.4×84.1 cm (A1) sheet of paper. Furthermore, all panellists were asked to write aroma descriptors of their own choice next to each wine.

The x- and y-coordinates of the positions of wines were measured. Sensory descriptors were collected for every wine sample and grouped according to similarity. From these descriptor groups, only those that had four or more entries for at least one wine sample were kept for further analysis. These groups were: 'clean/typical', 'fruity notes', 'floral notes', 'sweet notes', 'musty', 'ripe/apple/oxidized' and 'reductive notes'.

For Multiple Factor Analysis (MFA) the FactoMineR package (Lê et al., 2008) of the open source software R (version 3.1.1 (R Core Team, 2014)) was used. The x- and y-coordinates of each tasting sheet were defined as separate tables (blocks), while the descriptor groups were included into MFA as a categorical supplementary table as has been described by Pagès (2005b) and Perrin et al. (2008). The additional chemical data from FT-IR analysis and peak area values from important chromatogram segments (PARAFAC congruence loadings > 0.3) were also included into MFA as supplementary tables. More statistical background on MFA can be found elsewhere (Salkind, 2006; Escofier and Pages, 1994).

5.3 Results and discussion

The current study extends previous work from Chapter III and Chapter IV on the development of a non-targeted GC-MS screening method with experimental wine making using a full factorial experimental design and fast sensory profiling. This study provides new insights into main and interaction effects of the three factors grapevine age, yeast starter culture and must turbidity on the aroma composition of Riesling wines.

5.3.1 Fermentation performances

Each of the three studied factors had an impact on the fermentation process of the different treatments. Figure 5.1 shows the fermentation kinetics of all twelve treatments. All turbid musts fermented quicker and reached a high final degree of fermentation (less than $1\,\mathrm{g\,L^{-1}}$ residual sugar), while all fermentations of the clear musts proceeded slower and 'stuck' (fermentation has stopped before all the available sugar was metabolised) at the end of the fermentation. Final residual sugar contents of wines from clear musts of young and old grapevines were $17.9\,\mathrm{g\,L^{-1}}$ and $9.5\,\mathrm{g\,L^{-1}}$ for Oenoferm Klosterneuburg, $12.8\,\mathrm{g\,L^{-1}}$ and $7.3\,\mathrm{g\,L^{-1}}$ for Zymaflore X5, and $5.5\,\mathrm{g\,L^{-1}}$ and $4.3\,\mathrm{g\,L^{-1}}$ for EC 1118.

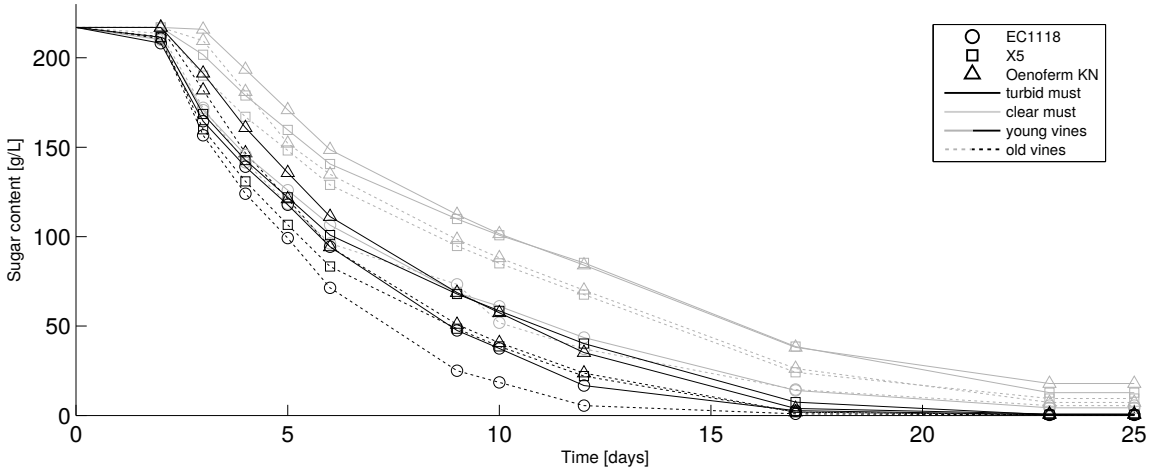


Figure 5.1: Fermentation kinetics. Wines are labeled as follows: OenfKN: Oenoferm Klosterneuburg, X5: Zymaflore X5, EC1118: EC1118.

Wines made from clear must of old grapevines had systematically lower residual sugar contents. All musts from old grapevines, irrespective of their turbidity, also showed slightly higher fermentation rates compared to the musts from younger grapevines, with the exception of the wine made from clear musts fermented with the yeast EC 1118. This treatment did not show any differences in fermentation kinetics between the musts from old and young grapevines. The starter culture EC 1118 also generally showed a better fermentation performance compared to the starter cultures

Zymaflore X5 and Oenoferm Klosterneuburg. These results were expected as EC 1118 is *S. bayanus*, which are known for a robust fermentation performance. The effects observed for the age of grapevines and the degree of must clarification on the fermentation are probably caused by different nutrition scenarios in the must. Moreover, solids in turbid musts provide a larger inner surface where yeast cells can adhere on which favours yeast activity/growth.

5.3.2 Non-targeted HS-SPME-GC-MS analysis

GC-MS analysis was performed using HS-SPME sample preparation due to the simplicity of this technique for wine analysis regarding full automation, speed and sensitivity. For HS-SPME a 100 µm PDMS fibre was used, as PDMS degradation products are easy to identify by means of siloxane fragments in their mass spectra. To facilitate a higher sample throughput a fast temperature ramp was applied to keep the GC runtime low. Lower chromatographic resolution was acceptable, as the non-targeted data analysis method used here takes the full mass dimension into consideration. The stability and reproducibility of the GC-MS system was monitored throughout the analysis period, which is particularly important when non-targeted analysis is applied. A fibre-blank injection and a hydro-alcoholic standard solution containing common wine volatiles were injected regularly after every eight sample analyses. This monitoring ensured the reproducibility of analyses. Absolute peak area values of the internal standard did not significantly (ANOVA, $\alpha = 0.05$,) differ among treatments, even though some wines had higher residual sugar contents. This fact made the use of an internal standard - although one was added to each sample as a precautionary measure - unnecessary.

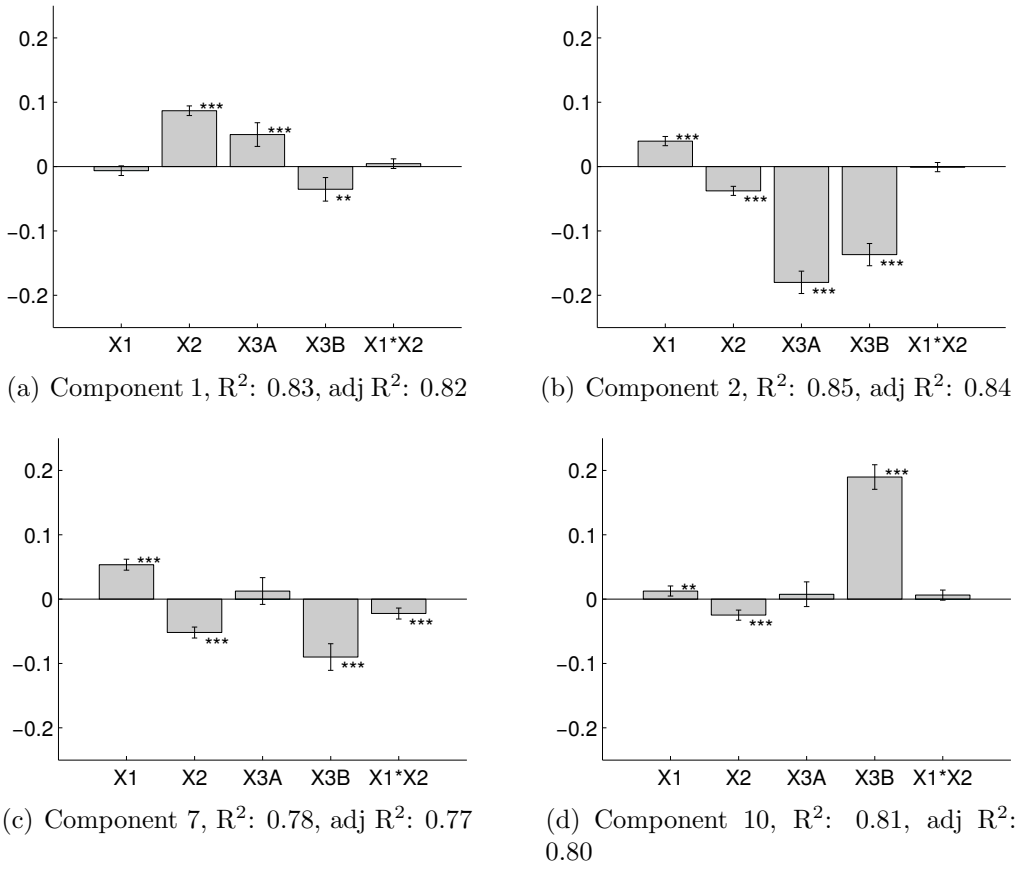
5.3.2.1 Data analysis GC-MS fingerprinting

The segmentation of chromatograms was done by examining overlays of all total ion chromatograms (TIC), while special care was taken that not too many peaks were included into each segment and no peak was allowed to shift into a neighbouring segment. In this manner 56 regions in the chromatograms were defined. Subsequently, every two neighbouring segments were combined to evaluate the impact of segment size resulting in one data set with 56 smaller segments and one with 28 larger segments. Mathematical transformation of the segmented chromatograms resulted in two three-way arrays with the dimensions $96 \times 96 \times 56$ and $96 \times 96 \times 28$. Here the first and second mode represent the wine samples (including replicates) and the third mode represents the chromatogram segments. Multiple models were built to find a PARAFAC model with an appropriate number of components. Each model was repeated 10 times to evaluate the stability and convergence time of each model. Moreover, the core consistency diagnostic (Bro and Kiers, 2003), residuals, and captured variance were examined. Outliers in the sample as well as in the segment mode were identified using Hotelling's T-Square statistics, by examining residuals and loadings, and were subsequently removed. The final three-way arrays after exclusion of outliers had the dimensions $95 \times 95 \times 49$ and $95 \times 95 \times 22$. The first two modes were mean centered and the last mode was scaled to unit variance using the nprocess.m function of the N-way toolbox. The PARAFAC models on both arrays revealed the same information on systematic differences among samples. The size of the segments had no influence on the quality of the PARAFAC results. Only the results of the PARAFAC model on the $95 \times 95 \times 22$ array is therefore discussed and represented in the following. A 14 component PARAFAC model gave the best interpretable results by explaining 89.8 % of the total variation in the dataset.

5.3.2.2 MLR and visual interpretation of PARAFAC components

In addition to visual examination of the PARAFAC loadings of the sample mode, multiple linear regression (MLR) was used as a complementary means to evaluate the effects of the three studied factors, as the results of the PARAFAC model can be interpreted as a projection of the raw data to a much lower dimensional space (projection to a few 'latent variables'). The PARAFAC loadings of the sample mode represent condensed information of the variation between samples. For each PARAFAC component a MLR model was therefore calculated using the PARAFAC loadings of the sample mode (duplicate injections were averaged) as the response variables (dependent variables) and the design matrix (Table 5.2) as independent variables. The MLR coefficients reflect a quantitative measure of a factor on the response variable. The higher a coefficient the higher its impact on the response variable. The coefficients for the two dummy variables representing the first and the second yeast starter cultures must be interpreted in relation to the third yeast starter culture, which is represented as zero (See coding in Table 5.2). For instance, if b_{3A} is positive and b_{3B} is negative, then yeast one gives the highest response, followed by yeast three and yeast two gives the lowest response.

Careful visual examination of the loadings of the sample mode of the PARAFAC model, absolute values and the significance of MLR coefficients revealed that four out of the 14 PARAFAC components, namely component one, two, seven and ten, contained information on systematic differences among the wines. The coefficients of the MLR models on the sample loadings of the PARAFAC component one, two, seven and ten and their significance levels based on Student t-test are represented in Figure 5.2. The remaining ten components represent unsystematic variations in the chromatograms and are not further discussed. In other words the data analysis approach applied here separates useful and non-essential information concerning the covariation among the chromatograms of all samples. Figures 5.3, 5.4 and 5.5 show the



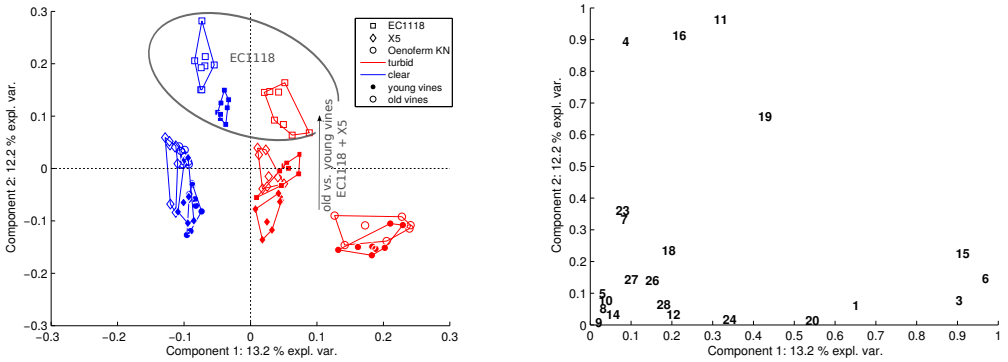
(a) Component 1, R^2: 0.83, adj R^2: 0.82

(b) Component 2, R^2: 0.85, adj R^2: 0.84

(c) Component 7, R^2: 0.78, adj R^2: 0.77

(d) Component 10, R^2: 0.81, adj R^2: 0.80

Figure 5.2: MLR coefficients according to the model postulated in Equation 5.1. Factor X_1: age of vines, factor X_2: must turbidity, factor X_3: yeast strain (see Table 5.1 for further details on the factorial design). Response variables are the PARAFAC loadings of the sample mode of each of the components. Significance is indicated as follows: $p > 0.05$ *, $p > 0.01$ **, $p > 0.001$ ***. Adjusted R^2 = adj R^2



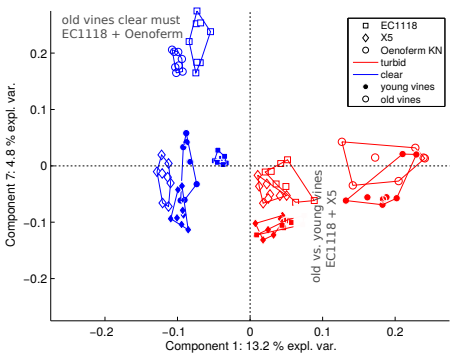
(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

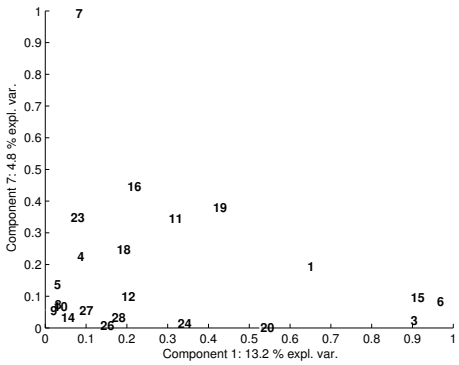
Figure 5.3: PARAFAC loadings: component one vs. two.

loadings of the important PARAFAC components. PARAFAC components two, seven and ten are all plotted against component one to facilitate an easier interpretation of the results. The congruence loadings of the segment mode (mode three) of the four important PARAFAC components represent the influence of a segment on the corresponding component (see Table 5.3). All segments with congruence loadings larger than 0.3 were considered to be 'week to strongly correlated' with the raw data, and therefore considered as important.

The coefficients of the MLR on the sample loadings of PARAFAC component one show that the factors must turbidity (X_2) and yeast strain (X_3) are significant (Figure 5.2(a)). The absolute size of the coefficients show that turbidity has the highest impact, while the effect of the yeast is slightly smaller. All coefficients, except of the interaction between the age of grapevine and must turbidity (X_1*X_2), of the MLR on the sample loadings of PARAFAC component two are highly significant (Figure 5.2(b)). The impact of the yeast starter cultures (X_3) is however four to five times higher than the impact of the age of grapevines (X_1) and must turbidity (X_2). The effects of the studied factors on the sample loadings of PARAFAC component seven is shown in Figure 5.2(c). The coefficients of X_1, X_2 and X_{3B} (age of grapevines, must turbidity and yeast Zymaflore X5) are highly significant and have equal total



(a) First mode (samples) loadings

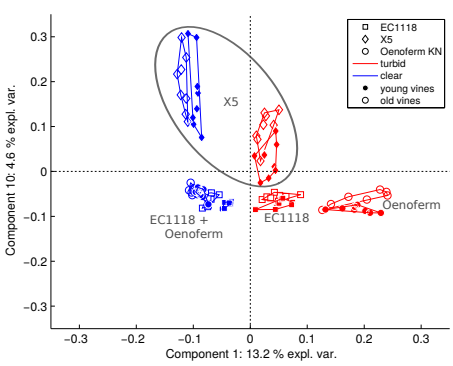


(b) Third mode (segments) congruence loadings

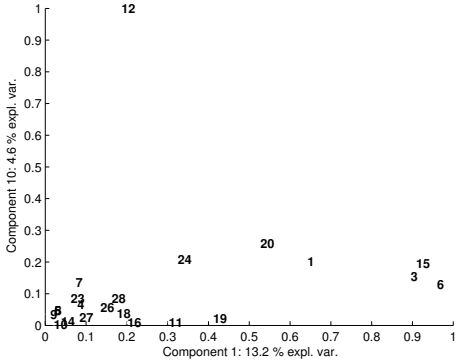
Figure 5.4: PARAFAC loadings: component one vs. seven.

values. Moreover, the interaction between the factors age of grapevines and must turbidity (X_1*X_2) is significant, albeit very small. The fourth MLR model on the sample loadings of component ten shows that the yeast starter culture Zymaflore X5 has a large impact in this case (Figure 5.2(d)), while all other factors have minimal effect, albeit the factors age of grapevines (X_1) and must turbidity (X_2) are significant. The examination of the MLR models showed clear main effects of the studied factors. The second order interaction effect between the factors age of grapevines and must turbidity (X_1*X_2) was negligibly small on all PARAFAC components. Coefficients for interaction effects between the yeast strain (a categorical factor with three factor levels) and the other two factors age of grapevines and turbidity cannot be calculated. A quantitative measure for these interactions can therefore not be provided. These dependencies of the factors can however be assessed by visual examination of the PARAFAC loadings.

What can be learned from the visual examination of the four important PARAFAC components will be discussed in greater detail in the following. Component one explaining 13.2 % of the variation in the dataset mainly reflects differences between the wines made from turbid and clear musts (Figure 5.3). The wines made from turbid must fermented with the yeast Oenoferm Klosterneuburg are separated from the other



(a) First mode (samples) loadings



(b) Third mode (segments) congruence loadings

Figure 5.5: PARAFAC loadings: component one vs. ten.

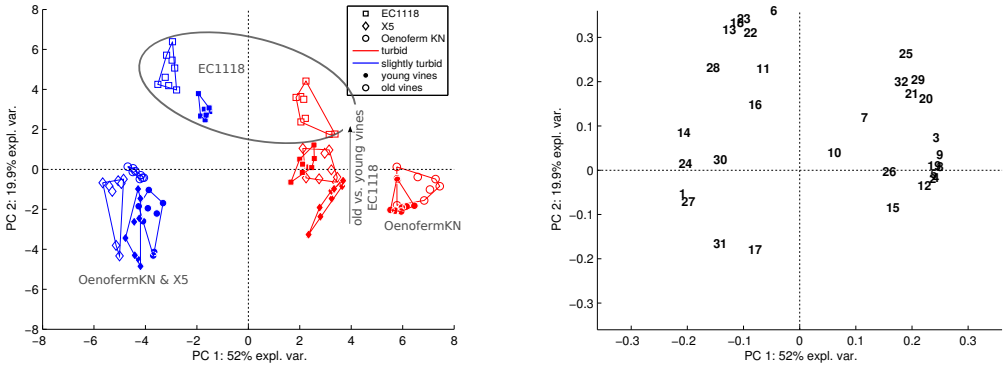
turbidly fermented wines by stronger correlation with this component. Component two accounts for 12.2 % of the variation in the dataset. The wines fermented with the yeast EC1118 made from clear must are positively correlated with component two, whereas the treatments from old grapevines correlate more with this component than the treatments from young vines. This effect of the age of grapevines can be observed for the turbidly fermented wines of the starter cultures EC1118 and Zymaflore X5. Wines made from clear must of old grapevines fermented with the yeasts EC1118 and Oenoferm Klosterneuburg are positively associated with component seven (4.8 % explained variance, Figure 5.4). This effect of grapevine age is also observed to a lesser extent by the wines obtained from turbid must fermented with the starter cultures Zymaflore X5 and EC1118 (Figure 5.4). Component ten, explaining 4.6 % of the variation in the dataset, shows differences between wines fermented with the starter culture Zymaflore X5 and all other wines (Figure 5.5). The wines made of clear must fermented with Zymaflore X5 tend to relate more to this component than the equivalent wines made of turbid must. Visual examination of the PARAFAC loadings clearly revealed that the volatile composition of the wines is effected by all studied factors (age of vines, turbidity and yeast strain) and that these factors influence each other.

Figures 5.3(b), 5.4(b) and 5.5(b) show the importance of each segment on the corresponding PARAFAC components. Segments 1, 3, 4, 6, 7, 11, 12, 15, 16, 19, 20, 23 and 24, which all have congruence loadings higher than 0.3 on the segment mode of the PARAFAC components one, two, seven and ten, were examined more closely to investigate the compounds responsible for the differences between groups of samples. A very conservative value of 0.3 was chosen here, which can be interpreted as a 'weak to strong correlation' with the raw data. The software package AMDIS was used to deconvolute coeluting peaks in each of the important segments. All peaks which showed significant (ANOVA, $\alpha = 5\,\%$) differences among treatments were tentatively identified by comparing deconvoluted mass spectra with the NIST 08 spectral library. Moreover, linear retention indices (LRI) were calculated using a homologous series of *n*-alkanes and compared with literature values to confirm tentative identifications. All compounds are summarized in Table 5.3.

5.3.2.3 PCA on deconvoluted peak areas

To verify the information obtained from the PARAFAC approach and to gain more detailed information on the compounds responsible for the discrimination between samples, a final PCA (with autoscaling) on the peak areas of all compounds was calculated. The first four principal components (PCs) reflect almost the same information as obtained from the PARAFAC approach regarding discrimination between groups of samples. Scores and loadings plots of all PCs (all plotted against PC1) are shown in Figures 5.6, 5.7 and 5.8.

In line with the first component of the PARAFAC model, the first PC, explaining 52.0 % of the total variance, separates treatments according to the degree of turbidity of the musts (Figure 5.6). PC1 correlates positively with the wines obtained from turbid musts as well as the branched alcohols isobutanol (compound 2), 2-methylbutanol (compound 3), isoamyl alcohol (compound 4), 2-phenylethanol



(a) scores (b) loadings

Figure 5.6: PCA scores and loadings: PC1 vs. PC2.

(compound 19), the branched fatty acid esters ethyl isobutyrate (compound 5), ethyl 2-methylbutanoate (compound 8), ethyl 3-methylbutanoate (compound 9), isobutyl hexanoate (compound 20), isopentyl hexanoate (compound 25), isoamyl octanoate (compound 29), isoamyl pentadecanoate (compound 32), and the two unknown compounds 12 and 26. Conversely, the wines made from clear musts correlate negatively with this PC. The acetate esters ethyl acetate (compound 1) and hexyl acetate (compound 14), the unsaturated ester ethyl 9-decenoate (compound 27), the sesquiterpene nerolidol (compound 30) and the unknown compound 24 show negative loadings on PC1. According to the information revealed from component 2 of the PARAFAC model, the second PC of the PCA (19.9 % explained variance) relates mainly to the wines fermented with the yeast EC1118 (Figure 5.6). The wines obtained from clear musts of old and young grapevines fermented with this yeast correlate highly with the fatty acid esters ethyl butanoate (compound 6), ethyl hexanoate (compound 13), ethyl octanoate (compound 23), methyl octanoate (compound 18), ethyl decanoate (compound 28), *m*-cymene (compound 16), the acetate ester isoamyl acetate and octanoic acid (compound 22). The wines made from turbid must of old grapevines fermented with the yeast EC1118 correlate with the esters from branched alcohols such as isobutyl acetate (compound 7), isobutyl hexanoate (compound 20), isopentyl

Table 5.3: Summary of all segments and their corresponding tentatively identified compounds showing high loadings (congruence loadings > 0.3) on PARAFAC components one, two, seven and ten.

	congruence loadings of PARAFAC component							
segment no.	1	2	7	10	no.	compound name[a]	LRI[b]	MS match
1	0.63				1	ethyl acetate	601	951
					2	isobutanol (2-methylpropanol)	613	911
3	0.89				3	2-methylbutan-1-ol	725	927
					4	isoamyl alcohol (3-methylbutanol)	721	945
					5	ethyl isobutyrate (ethyl 2-methylpropanoate)	748	887
4		0.89			6	ethyl butanoate	796	940
					7	isobutyl acetate (2-methylpropyl acetate)	765	809
6	0.95				8	ethyl 2-methylbutanoate	847	871
					9	ethyl 3-methylbutanoate	849	812
7		0.33	0.99		10	hexanol	866	963
					11	isoamyl acetate (3-methylbutyl acetate)	873	977
					12	unknown m/z(%) = 104(100), 103(45), 151(40), 78(31), 51(14), 207(10), 105(7)	899	
11	0.30	0.96	0.35		13	ethyl hexanoate	999	936
					14	hexyl acetate	1011	968

Table 5.3 – continued

	congruence loadings of PARAFAC component							
segment	1	2	7	10	no.	compound name[a]	LRI[b]	MS match
12				1.00	15	unknown m/z(%) = 121(100), 93(85), 136(74), 91(44), 105(17), 107(11), 103(10)	1030	863
					16	*m*-cymene (1-isopropyl-3-methylbenzene)	1038	811
					17	ethyl 2-hexenoate (ethyl (E)-hex-2-enoate)	1042	901
15	0.90				18	methyl octanoate	1119	952
					19	2-phenylethanol	1130	958
					20	isobutyl hexanoate (2-methylpropyl hexanoate)	1145	845
16		0.91	0.45		21	unknown m/z(%) = 101(100), 129(76), 128(19), 102(13), 55(10), 73(8)	1171	
					22	octanoic acid	1174	906
					23	ethyl octanoate	1196	931
					24	unknown m/z(%) = 59(100), 93(95), 121(80), 136(80), 81(58), 43(34), 92(32), 95(20)	1215	
17[c]					25	isopentyl hexanoate (3-methylbutyl hexanoate)	1245	882
					26	unknown m/z(%) =99(100), 163(38), 117(10), 105(9)	1248	
19	0.41	0.66	0.38		27	ethyl 9-decenoate (ethyl dec-9-enoate)	1382	888

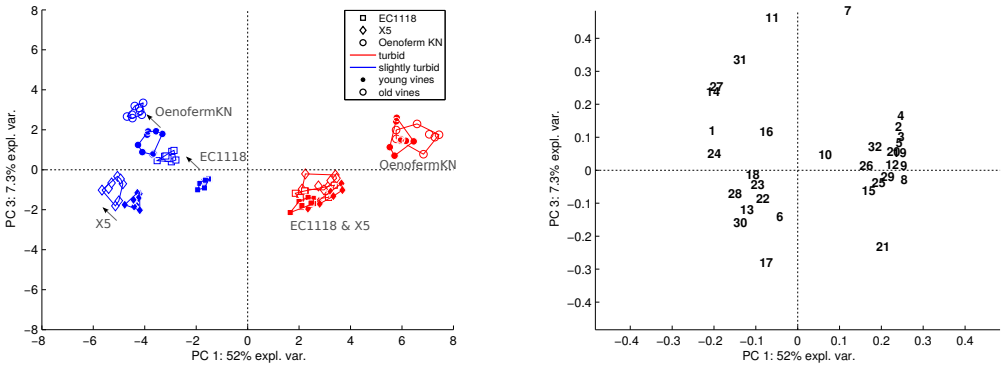
Table 5.3 – continued

segment	congruence loadings of PARAFAC component 1	2	7	10	no.	compound name[a]	LRI[b]	MS match
					28	ethyl decanoate	1393	929
20	0.53				29	isoamyl octanoate (3-methylbutyl octanoate)	1441	916
23		0.36	0.35		30	nerolidol (3,7,11-trimethyl-1,6,10-dodecatrien-3-ol)	1572	834
					31	unknown m/z(%) = 183(100), 57(63), 43(58), 71(27), 85(27), 55(21), 145(19), 95(16)	1578	
24	0.33				32	isoamyl pentadecanoate (3-methylbutyl pentadecanoate)	1641	818

[a]For each segment only compounds showing significantly different peak area values between treatments are listed.

[b]Experimentally determined linear retention indices.

[c]Segment 17 was excluded from the PARAFAC model due to interfering signal in this segment.



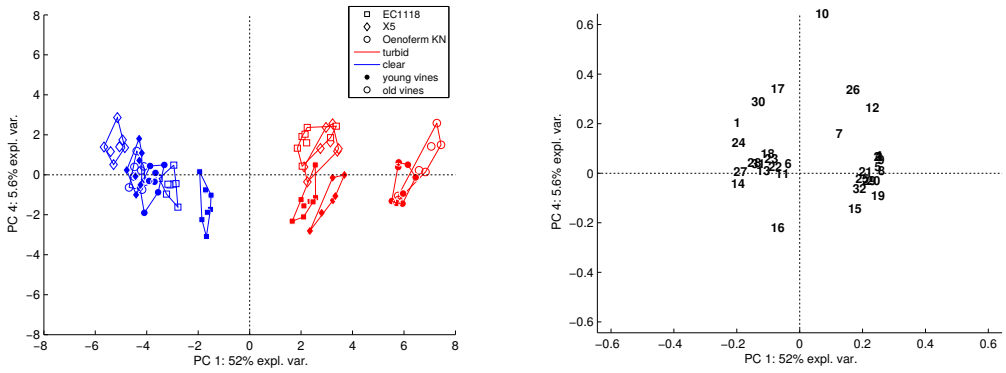
(a) scores (b) loadings

Figure 5.7: PCA scores and loadings: PC1 vs. PC3.

hexanoate (compound 25), isoamyl octanoate (compound 29), isoamyl pentanoate (compound 32) and the unknown compound 21, whereas the equivalent wines made from young grapevines grouped together with the wines from turbid must fermented with the other two starter cultures.

The differences between the wines from musts from old and young grapevines which could be observed in PARAFAC component 7 are associated with PC3 and PC4 (7.4 % and 5.6 % explained variance, respectively; Figure 5.7 and 5.8). The wines made from clear musts obtained from old grapevines fermented with the yeasts EC1118 and Oenoferm Klosterneuburg correlate positively with hexyl acetate (compound 14), ethyl 9-decenoate (compound 27) and the unknown compound 31. These wines also correlate negatively with the unknown compound 21.

Overall the results of the PCA based on chromatographic peak areas are comparable to those of the PARAFAC approach. However, information on the differentiation of the wines made with the starter culture Zymaflore X5 which could be observed on PARAFAC component 10 could not be extracted from PCA data, although all wines made with this yeast have significantly higher (2-3 fold) levels of the compound ethyl 2-hexenoate (compound 17), as shown in the boxplot in Figure 5.9. A similar scenario has been described in Chatper III, where the PARAFAC approach used here for pro-



(a) scores (b) loadings

Figure 5.8: PCA scores and loadings: PC1 vs. PC4.

cessing of GC-MS chromatograms was superior to PCA performed on deconvoluted peak area values. Ethyl 2-hexenoate (compound 17) is assumed to be a varietal ester formed from precursors located in the grape skins (Antalick, 2010). Levels of this ester can be effected among others by yeast (Liang et al., 2013) and leaf removal in the vineyard (Šuklje et al., 2014).

5.3.3 Merging of chemical and sensory data

MFA on the x- and y-coordinates of wines from partial projective mapping revealed systematic differences among samples similar to those obtained from the GC-MS fingerprinting of volatiles. In order to establish links between chemical and sensory data, the citation frequencies of descriptor groups (how often a descriptor was mentioned for a wine), the area values of deconvoluted peaks and additional chemical data from FT-IR analysis were integrated into MFA as supplementary tables. After outlier removal, 14 of the 18 tasters were included in the MFA. The representation of wines is given in Figure 5.2(a). The wines from turbid musts fermented with the yeast Oenoferm Klosterneuburg (OenfKN_O_T and OenfKN_Y_T) correlate with the aroma descriptor group 'musty'. The projected chromatographic data revealed correlations of this descriptor group with the branched alcohols isobutanol (compound 2),

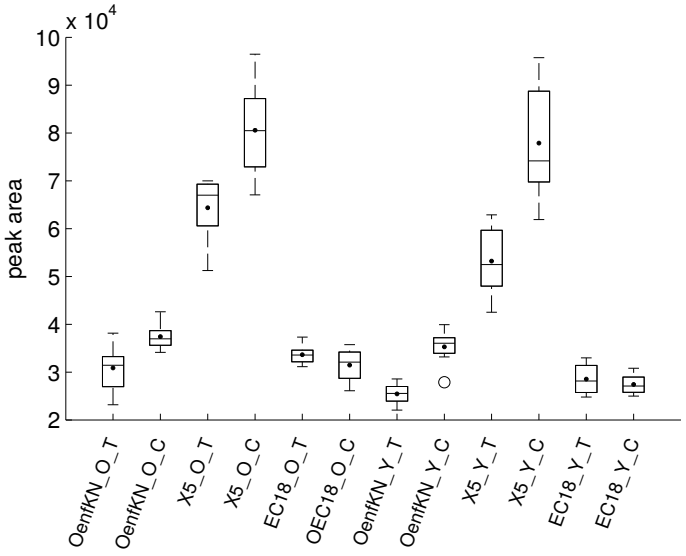
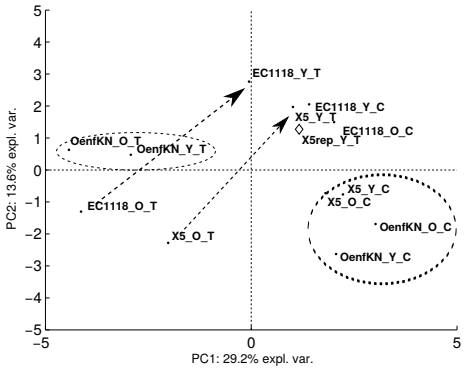


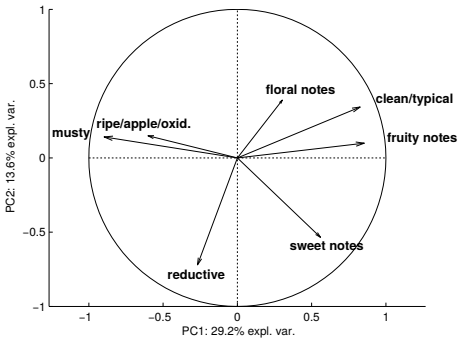
Figure 5.9: Boxplot of ethyl 2-hexenoate peak areas for each of the experimental wines. Wines are labeled as follows: OenfKN: Oenoferm Klosterneuburg, X5: Zymaflore X5, EC1118: EC1118, O: old vines, Y: young grapevines, T: turbid must, C: clear must.

2-methyl butanol (compound 3), isoamyl alcohol (compound 4) and 2-phenylethanol (compound 19), the branched fatty acid esters ethyl isobutyrate (compound 5), ethyl 2-methylbutanoate (compound 8), ethyl 3-methylbutanoate (compound 9), isobutyl hexanoate (compound 20), isopentyl hexanoate (compound 25), isoamyl octanoate (compound 29) and isoamyl pentadecanoate (compound 32), as well as the two unknown compounds 12 and 26. The same compounds also reflected the differentiation of these wines in the PCA performed on the autoscaled peak table, as shown in Figure 5.6. Negative sensory effects of highly turbid musts have been often described in literature (Singleton et al., 1975; Groat and Ough, 1978; Houtman and Du Plessis, 1981; Losada et al., 2011). All wines made from clear musts but also the wine X5_Y_T (turbid must from young grapevines fermented with Zymaflore X5) correlate negatively with above-mentioned compounds and positively with the sensory descriptor groups 'clear/typical' and 'fruity notes'.

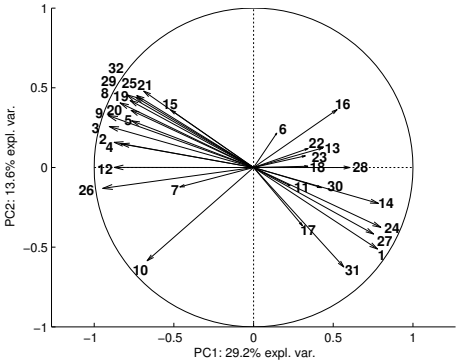
The samples EC1118_O_T and X5_O_T (turbid musts from old vines, fermented with Zymaflore X5 and EC1118) are the only treatments where the grapevine age



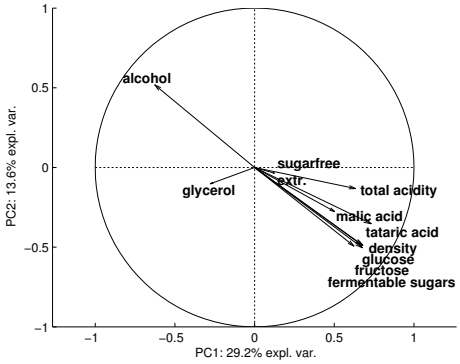
(a) Representation of wines (common factor scores)



(b) Representation of sensory descriptors (correlations)



(c) Representation of volatile compounds (correlations)



(d) Representation of additional chemical data (correlations)

Figure 5.10: Results of MFA of the partial projective mapping (orthonasal evaluation only). Wines in (a) are labeled as follows: OenfKN: Oenoferm Klosterneuburg, X5: Zymaflore X5, EC18: EC1118, O: old vines, Y: young wines, T: turbid must, C: clear must. Frequencies of the aroma descriptor groups of the free choice profiling (b) were included as categorical supplementary variables. Autoscaled peak areas (c, numbers correspond to compounds in Table 5.3) and data from FT-IR (d) were included as continuous supplementary variables.

affected on the sensory impression. Wines from old and young grapevines of all other treatments were rated similarly. An effect of the age of grapevines from turbid musts fermented with these two yeast starter cultures is also reflected on PARAFAC component 7 of the GC-MS fingerprinting data. Segment 7 has high congruence loadings on PARAFAC component 7 (Figure 5.4). Accordingly, the representation of volatile compounds in MFA Figure 5.10(c) shows that hexanol (compound 10 in the chromatogram segment 7) correlates with the wines EC1118_O_T and X5_O_T. Moreover, the representation of sensory attributes displayed in Figure 5.10(b) illustrates correlation of the sensory descriptor group 'reductive notes' with these wines and the compound hexanol (compound 10). It is well known that reductive notes are caused by sulphur containing compounds such as H_2S and/or mercaptans. A direct causal link between reductive notes and higher levels of hexanol (compound no. 10) is therefore very unlikely. A possible negative sensory impact of this compound on wine aroma has however been described before. Marais and Pool found a negative correlation between the intensity of the young wine bouquet and levels of hexanol, 2-methylpropanol and 3-methylpropanol in Riesling and other varieties (Marais and Pool, 1980). Moreover, Rankine and Pococx reported in their sensory study that higher concentrations of hexanol tended to give wine a foreign aroma, which they regarded as a reduction in quality (Rankine and Pococx, 1969). The wines made of clear musts from old and young grapevines fermented with Zymaflore X5 and Oenoferm Klosterneuburg (OenfKN_O_C, OenfKN_Y_C, X5_O_C and X5_Y_C) form a separate group correlating with the descriptor group 'sweet/honey' (Figure 5.10(a) and 5.10(b)). Glucose, fructose, fermentable sugar, and density also correlate with these wines as can be seen in the representation of the additional chemical data in Figure 5.10(d). The higher sugar levels of these wines result from stuck fermentation close to the end (Figure 5.1). Interestingly, although all samples were only orthonasally evaluated, the wines with residual sugar were grouped together. The acetate esters

ethyl acetate (compound 1) and hexyl acetate (compound 14), the unsaturated ethyl ester ethyl 9-decenoate and two unknown compounds 24 and 31 correlate with these wines (Figure 5.10(c)). The same compounds also correlated with these samples on PC1 and PC2 of the PCA performed on the autoscaled peak table (Figure 5.6).

The results presented here reveal that the decrease in overall wine quality observed in sensory analysis is not only dependant on the turbidity of musts, but also depends on the yeast strain and the composition of the must (musts from old grapevines vs. musts from young vines). Effects of the studied factors (grapevine age, must turbidity and yeast strain) and how they influence each other highlight the benefits of the multifactorial approach used in this study. The results show that similar information on the grouping of wines were obtained from GC-MS fingerprinting and partial projective mapping. By incorporating area values of the peaks that contributed to differences among samples in the PARAFAC model into MFA, correlations of the compounds with sensory descriptor groups could be found. However, correlations between volatile compounds and aroma descriptor groups have to be interpreted with some caution, as correlations do not necessarily imply causality.

5.4 Conclusions

To obtain a better understanding of the importance of viticutural and oenological factors and their interactions on the composition of wine aroma compounds, experimental wine making in combination with sensory and chromatographic analysis is essential. In this study, non-targeted HS-SPME-GC-MS fingerprinting and partial projective mapping with free choice descriptor profiling were combined with full-factorial design of experimental wine making to allow an in-depth study of the impact of the age of vines, must turbidity and yeast starter culture on the volatile composition and the aroma of Riesling wines. The applied GC-MS fingerprinting approach (approach 1, Section 3.4), including segmentation and transformation of chro-

matograms combined with PARAFAC modelling revealed differences between wine samples. Not all differences between samples discovered by the PARAFAC approach could be fully reproduced by simple PCA on autoscaled peak table data (deconvoluted peak areas), which points out the benefit of the PARAFAC approach over only PCA on peak tables. The use of full factorial experimental design with visual examination of the results of the fingerprinting approach and MLR revealed main effects and interaction effects between studied factors on groups of compounds. The integration of information from fast GC-MS screening of volatiles and rapid sensory profiling by means of the multi-block PCA method MFA facilitates the correlation of compounds with sensory descriptor groups and wine samples. These correlations have to be very carefully interpreted as a correlation does not necessarily imply causality. Main and interaction effects of the factors vine age, must turbidity and yeast strain on the aroma of Riesling wines could be shown. For instance, the sensory impression of wines made from turbid musts of old and young grapevines were rated differently for two of the three yeast starter cultures. Different yeast starter cultures reacted differently to must turbidity, and this effect even depended on the composition of the must (must from old grapevines vs. must from young vines). The results presented herein emphasise the need for multifactorial approaches including multivariate statistics to study the impact of oenological and viticultural factors on wine aroma. The discovered effects are very likely to be influenced by even more factors such as grape variety, vintage, clones, location of the site, and others. Full factorial designs however quickly become too big and complex the more factors and factor levels are included. The application of screening designs prior to full factorial examination of influencing factors could be a solution to this problem in future studies.

CHAPTER VI

General conclusions

The primary goal of this study was the development and application of a new data analysis approach for non-targeted gas chromatography mass spectrometry (GC-MS) fingerprinting data of wine volatiles with a special focus on the avoidance of retention time correction between samples and feature selection. Matrix algebra and chemometrics were used for mathematical transformations and modelling of two-dimensional GC-MS chromatograms of multiple samples. Moreover, merging of data from non-targeted GC-MS fingerprinting of volatiles and fast sensory profiling was another focus of this study.

In the first chapter, general background on targeted and non-targeted chromatographic analysis with an emphasis on chemometrical modelling of chromatographic data is provided. Furthermore, the composition of wine aroma in terms of the volatile composition of wine and the use of gas chromatography in wine analysis is reviewed. A short introduction into rapid sensory profiling of wine is also given.

The development of two new chemometric approaches for non-targeted GC-MS data is presented in Chapter three. A major drawback of conventional data analysis approaches is the necessity of retention time alignment of peaks between samples. Some existing chemometric approaches use multivariate (or multi-way) models which take peak shifts between samples into account to deconvolute predefined chro-

matogram segments resulting in deconvoluted peak profiles for each segment and sample. Building and evaluating one multivariate model for each chromatogram segment for all (or even each) sample is however very time consuming. The two approaches (algorithms) described in Chapter three take peak shifts among samples into account and are applied to the entire chromatograms (all predefined chromatogram segments) of all samples. The results reveal information on systematic differences among samples and the importance of chromatogram segments contribution to differences among samples. Only these important chromatogram segments containing information on differences among samples can then subsequently be deconvoluted, if further information on the chemical compounds in these segment is needed. This represents a vast saving in time as only a small number of important segments has to be deconvoluted.

Both approaches use segmentation of the chromatograms and subsequent transformation of the two-dimensional chromatogram segments of each sample (*mass spectral profile* $\times$ *elution profile*) into sums of squares and cross product matrices (SSCP; *mass spectral profile* $\times$ *mass spectral profile*). The sums of squares are a measure of variation within a mass channel, whereas the cross products are a measure of covariation between two mass channels. Note, that SSCP matrices are similar to variance-covariance matrices. The SCCP matrices of chromatogram segments with peaks of the same concentration in different samples remain constant, even when the location (retention time) of the peaks are different among samples. Besides the described segmentation and transformation, approach one includes further mathematical rearrangements resulting in a three-way array which can be decomposed using parallel factor analysis (PARAFAC). Visual examination of the PARAFAC loadings reveals sample groupings and important segments responsible for the groupings can be identified. Approach two is also based on segmentation of the chromatograms. Each segment is automatically decomposed using singular value decomposition (SVD), which is an eigenvalue decomposition of the SSCP matrix of the chromatogram segment.

Only the first singular value (or values) are used for further PCA analysis. Similar to approach one, PCA scores show groupings of the samples, while the loadings provide information on segments responsible for the grouping. Based on the results of approach one and approach two, important segments responsible for the grouping between samples, can be further deconvoluted, if more detailed information on the compounds in these segments is needed. The here developed approaches can also be considered as a segment selection tools for the deconvolution of chromatogram segments, as the number of segments for deconvolution is largely reduced compared to the deconvolution of all segments of a chromatogram.

The two approaches have been tested on an artificial data set and on a real HS-SPME-GC-MS data set of wines fermented with different yeast and malolactic fermentation scenarios. The results were compared to each other and validated with a reference method (PARAFAC2 deconvolution of all chromatogram segments with subsequent PCA). Both approaches are suitable for finding systematic differences among samples. The PARAFAC model in approach one is more difficult to model, whereas in approach two only SVD (and PCA) is utilized. Approach one, however, reveals more structure in the data then approach two. Moreover, even the PCA results from deconvoluted peaks of all segments (reference method) showed less information on differences among samples then the results of approach 1. Approach 1 is therefore a fast and more effective alternative to conventional data analysis methods. The suitability of approach 1 for large data sets, such as metabolomics data, where samples are analysed in multiple sequences and contain therefore more shifting peaks has to be still investigated. The only manual tasks of approach 1 are the segmentation of chromatograms and PARAFAC modelling. Automated segmentation would be however essential for the analysis of LC-MS data, where a visual segmentation of the chromatograms would not be possible due to the much bigger mass-to-charge range (50 - 4000 u) compared to GC-MS data (30 - 500 u). Approach 1 has been applied

in the further course of the thesis to study relevant topics in wine research.

In Chapter four, approach one was used to study the effect of different malolactic fermentation scenarios on the volatile composition of a fresh, fruity Pinotage style and a full bodied Pinotage style. Moreover, sensory evaluation of the wines was carried out using the rapid descriptive method projective mapping (similiar to napping) with free choice descriptor profiling (Ultra Flash Profiling). Merging of the results of GC-MS fingerprinting and perceptual mapping by means of multiple factor analysis (MFA) provided a comprehensive integrated overview of the volatile composition and the sensory expression of the wines. Correlations of volatile compounds and sensory attributes were found. A high correlation was found between the attribute 'fruitiness' and many esters, which are well known for their contribution to fruity notes in wine. Furthermore, one wine mainly described with the descriptor 'vegetal/herbaceous' and correlated with higher concentrations of an unknown compound with an terpenoid-like mass spectra. The presented method was proven to be a fast and powerful tool to obtain a broad overview on the sensory characteristics and the volatile composition of experimental wines.

In Chapter five, the second application also involved the combined data evaluation of the perceptual mapping data and the GC-MS fingerprinting data obtained from approach one for a set of Riesling experimental wines. In this chapter, the experimental wine making was done in a full factorial design to allow in-depth study of the main effects and interaction effects of the viticultural factor grapevine age and the oenological factors yeast starter culture and musts turbidity on the aroma of the studied wines. Main and interaction effects of all factors on the aroma of Riesling wines could be shown. For instance, the sensory impression of wines made from turbid musts of old and young grapevines were rated differently for two of the three yeast starter cultures. Different yeast starter cultures reacted differently to must turbidity, and this effect depended on the composition of the must (must from old grapevines

vs. must from young vines). The multifactorial strategy used in this chapter shows how the effect of factors can correlate with each other, emphasising the importance and necessity of studying several possible factors at the same time.

Several general conclusions may be drawn from the results presented in this thesis. Non-targeted GC-MS fingerprinting of wine result in a more holistic view on the composition of wine volatiles compared to targeted methods, which are always focused on a certain set of *a priori* known and identified compounds. Matrix algebra and advanced chemometric modelling are powerful tools for alternative approaches to non-targeted chromatographic data analysis. Problems concerning feature selection and retention time correction when using conventional approaches can be avoided by applying mathematical transformations on the raw data points of the chromatograms and subsequent modelling. The data analysis approaches presented here offer a useful alternative to conventional methods. The development of such approaches require an 'out of the box' thinking, considering chromatograms as signals from an instrument and not as a sequence of peaks which 'have to' be integrated to obtain useful data. Programming skills are however necessary to implement algorithms.

Non-targeted GC-MS fingerprinting of wine provides comprehensive analytical data on the composition of the analysed wines. Drawing conclusions from the volatile composition of a wine to its sensory properties is not possible. The merging of data from sensory analysis and analysis of volatiles is therefore important. The presented possibility of combining data from the developed approach herein and projective mapping offers an effective tool to comprehensively study wine aroma by obtaining correlations between aroma descriptors and volatile compounds. The extension of this strategy to multifactorial experimental wine making contributed to significant new information regarding the main and interaction effects of grapevine age, yeast starter cultures and must turbidity on the aroma of Riesling wines.

The here developed data analysis approaches could be further adapted to other

analytical techniques such as comprehensive two dimensional gas chromatography coupled to mass spectroscopy (GC×GC-MS) and liquid chromatography mass spectrometry (LC-MS). Moreover, the implementation into a software package including a graphical user interface would make this data analysis approach accessible to analytical scientist without programming experience. Considering more applications in viticultual and oenological studies, the usage of more advanced experimental designs or the combination of screening designs and full factorial designs could facilitate the inclusion of more oenological and viticultural factors into aroma studies.

APPENDICES

APPENDIX A

Schematic representation of approach 1

The following Figure A.1 shows a schematic representation of approach 1 which is described in Section 3.4. MATLAB codes for approach 1 and approach 2 are provided in the Appendices C and D, respectively.

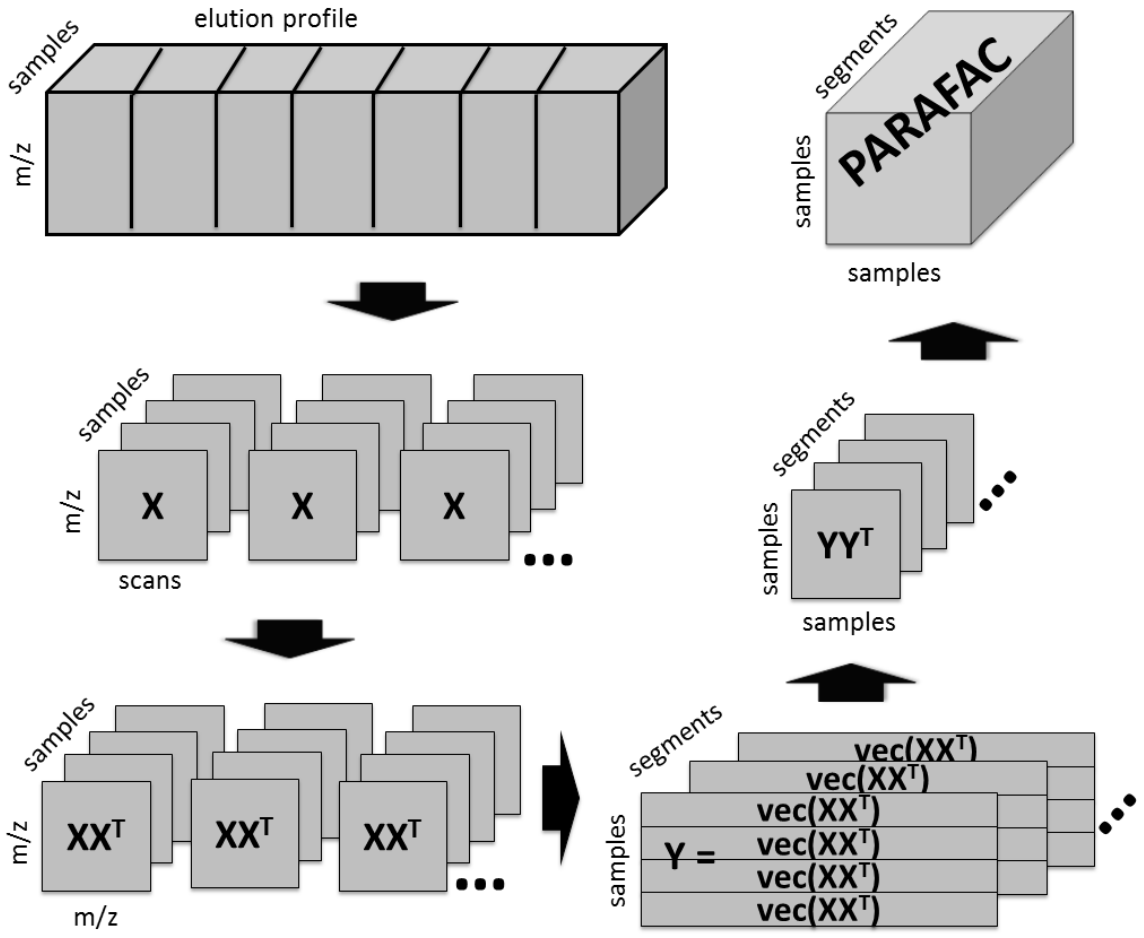


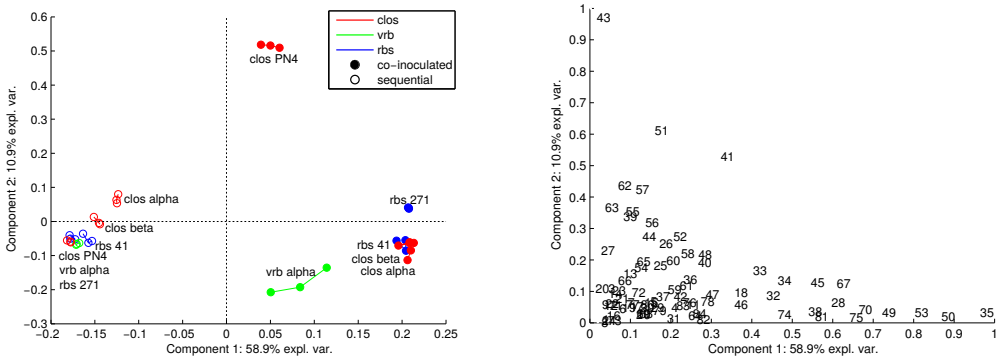
Figure A.1: Schematic representation of approach 1. Matrix indices were omited. Note that only the upper triangular matrices of all XX^T are used.

APPENDIX B

Approach 1 with class centroid centering and scaling to intra-class variance applied to the compilation matrices Y^k (Equation 3.9 in Section 3.6)

In the following the results of the application of approach 1 to the HS-SPME-GC-MS analysis of the Cabernet Sauvignon wines (Section 3.6), where class centroid centering and scaling to intra-class variance (12 classes, one for each treatment) was applied to the compilation matrices Y^k (Equation 3.9) are shown. A 9 component PARAFAC model explaining 98.1 % of variation in the dataset was obtained. Note that both the final three way array $\underline{Z}$ were centered across sample modes (mode one and two) and scaled to unit variance within the segment mode (mode 3). In general very similar information on the grouping of samples is obtained compared with the PARAFAC model on the unscaled compilation matrices and the results of approach 2. Component one reflects the differences between co-inoculated and sequentially inoculated wines (Figure B.1), component two the differences of the wine co-inoculated with the yeast/bacteria combination Lalvin Clos/Lalvin PN4 (Figure B.1), component four the differences of the sequentially inoculated wines fermented

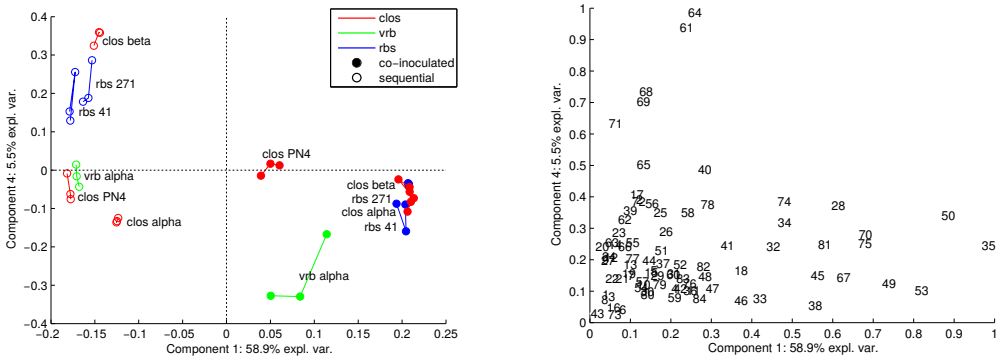
with the yeast starter culture Uvaferm RBS and the wine sequentially inoculated with the yeast/bacteria combination Lalvin Clos/Enoferm Beta (Figure B.2), component five the difference of the wine sequentially inoculated with the yeast/bacteria combination Lalvin Clos/Enoferm Alpha (Figure B.3), component seven the differences of all wines fermented with the yeast starter culture Uvaferm RBS (Figure B.4) and component nine the differences of all wines fermented with the yeast starter culture Uvaferm VRB (Figure B.4). All other PARAFAC components represented non-systematic variation between samples.



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

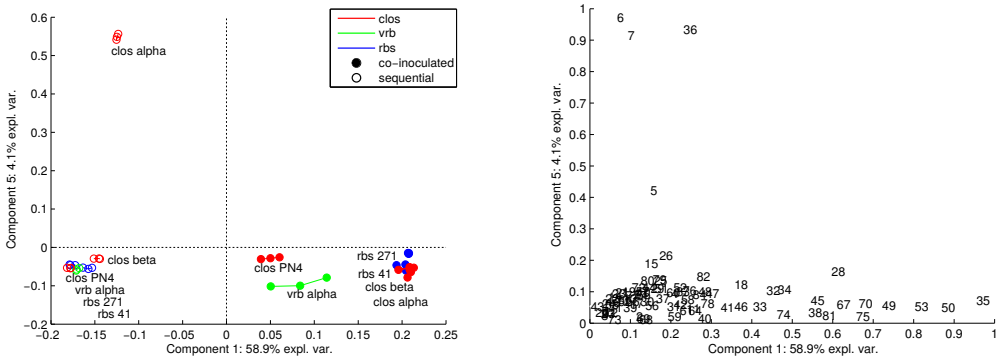
Figure B.1: Loadings plots of PARAFAC components one vs. two, where class centroid centering and scaling to intra-class variance was applied to the compilation matrix Y^k (Equation 3.9); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

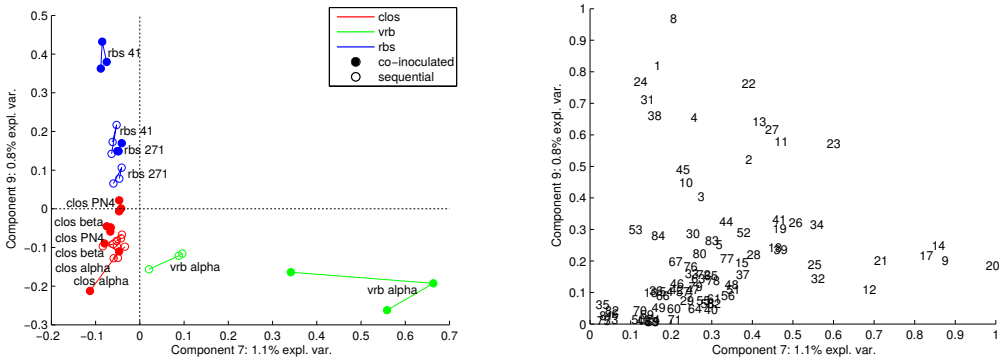
Figure B.2: Loadings plots of PARAFAC components one vs. four, where class centroid centering and scaling to intra-class variance was applied to the compilation matrix Y^k (Equation 3.9); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure B.3: Loadings plots of PARAFAC components one vs. five, where class centroid centering and scaling to intra-class variance was applied to the compilation matrix Y^k (Equation 3.9); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).



(a) First mode (samples) loadings

(b) Third mode (segments) congruence loadings

Figure B.4: Loadings plots of PARAFAC components seven vs. nine, where class centroid centering and scaling to intraclass variance was applied to the compilation matrix Y^k (Equation 3.9); Yeast starter cultures: Lalvin Clos (clos), Uvaferm RBS (rbs), Uvaferm VRB (vrb); Lactic acid bacteria starter cultures: Enoferm Alpha (alpha), Enoferm Beta (beta), Lalvin PN4 (PN4), Lalvin VP41 (41) and O-Mega (271).

APPENDIX C

MATLAB code approach 1

```
%% Load data
clear, clc
cd C:\directory\directory\directory\
load yourdata.mat % load data
% chromatograms arranged in 3-way array 'cube' (scans x mz x samples)
% classes in vector 'classes' eg. classes=[1 1 1 1 1 1 2 2 2 2 3 3 3 3]
% samplenames in cell 'samplenames'
%eg. samplenames={ 'sample1' 'sample2' 'sample3'  ... }
%% calculate TICs all samples and store in PLS-Toolbox dataset

figure
tics=dataset(squeeze(sum(cube,2)));
tics.class{2}=classes;  %assign classes
tics.label{2}=samplenames;  %assign sample names
plotgui(tics)            %plot TICS (PLS-Toolbox function)

%use data cursor to define segments, save them to workspace
%('cursor_info') and save in file 'cursor_info_section.mat'

% plot segment borders
% pause
% hold on
% vline(segments(:,1))
% text(round(mean(segments,2)),repmat(0,1,size(segments,1)), ...
%       num2str([1:size(segments,1)]'))
%% load segmentation and transform to variable 'segments'
%eg. segments=[1 50; 51  92; 93 151 ...  ]
load('cursor_info_section.mat'); %
sec=[];
for i=1:size(cursor_info,2)
    x=cursor_info(i).Position;
```

```
    x=x(1);
    sec(i)=x;
end
sec=sort(sec); sec=[1 sec size(cube,1)]
segments=[]
for i=1:size(sec,2)-1
    segments(i,1)=sec(i)
    segments(i,2)=sec(i+1)
end
%
%% transformations

% kick out segments eg. segmentsout=[1:16 17:50]
included_segments=[1:size(segments,1)]
included_segments=included_segments(segmentsout)

YY=[]
for i= 1:size(included_segments,2)
    disp(['section: ' num2str(included_segments(i)) ' from ' ...
        num2str(size(included_segments,2)) ] )
    Y=[];
    for c=1:size(cube,3)
        X=squeeze(cube(segments(included_segments(i),1): ...
            segments(included_segments(i),2),:,c))';
        %remove offset if necessary
        %         m=repmat(min(X'),size(X,2),1);
        %         Xm=X'-m;
        %         X=Xm';
        cross=X*X'; % cross product
        Y(c,:)=cross(:)'; %compilation matrix for each segment
    end
    YY=Y*Y';%cross product
    %optional scaling of the compilation matrix
    % XXX=dataset(XXX);
    % XXX.class{1}=classes;
    % scaling
    % [datap,sp] = preprocess('calibrate','classcentroidscale',XXX);
    % Y=datap.data*datap.data';%cross product

    threeway(i,:,:)=YY;
end

%% PARAFAC
%center and scale
Xmultiway=nprocess(threeway, [0 1 1], [1 0 0]);
Xmultiway=dataset(Xmultiway);
Xmultiway.label{1}=num2str(included_segments')
Xmultiway.axisscale{1}=included_segments;
Xmultiway.class{2}=classes;
Xmultiway.label{2}=samplenames;

numcomps= 10;
model=parafac(Xmultiway,numcomps); % PLS-Toolbox function
```

APPENDIX D

MATLAB code approach 2

```
%% Load data
clear, clc
cd C:\directory\directory\directory\
load yourdata.mat % load data
% chromatograms arranged in 3-way array 'cube' (scans x mz x samples)
% classes in vector 'classes' eg. classes=[1 1 1 1 1 1 2 2 2 2 3 3 3 3]
% samplenames in cell 'samplenames'
%eg. samplenames={ 'sample1' 'sample2' 'sample3'  ... }
%% calculate TICs all samples and store in PLS-Toolbox dataset
figure
tics=dataset(squeeze(sum(cube,2)));
tics.class{2}=classes;  %assign classes
tics.label{2}=samplenames;  %assign sample names
plotgui(tics)            %plot TICS (PLS-Toolbox function)

%use data cursor to define segments, save them to workspace
%('cursor_info') and save in file 'cursor_info_section.mat'

% plot segment borders
% pause
% hold on
% vline(segments(:,1))
% text(round(mean(segments,2)),repmat(0,1,size(segments,1)), ...
%       num2str([1:size(segments,1)]'))
%% load segmentation and transform to variable 'segments'
%eg. segments=[1 50; 51  92; 93 151 ...  ]
load('cursor_info_section.mat'); %
sec=[];
for i=1:size(cursor_info,2)
    x=cursor_info(i).Position;
    x=x(1);
```

```
    sec(i)=x;
end
sec=sort(sec); sec=[1 sec size(cube,1)]
segments=[]
for i=1:size(sec,2)-1
    segments(i,1)=sec(i)
    segments(i,2)=sec(i+1)
end
%
%% transformations
% kick out segments eg. segmentsout=[1:16 17:50]
included_segments=[1:size(segments,1)]
included_segments=included_segments(segmentsout)

numberofS=2;    % number how many singular values should be kept

allSs=[];
allSs_final=[];
for i= 1:size(included_segments,2)
    disp(['section: ' num2str(included_segments(i)) ' from ' ...
        num2str(size(included_segments,2)) ] )
    allS=[];
    for c=1:size(cube,3)
        %X is ith segment, of cth sample
        X=squeeze(cube(segments(included_segments(i),1): ...
            segments(included_segments(i),2),:,c))';
        %m=repmat(min(X'),size(X,2),1); % remove offset if necessary
        %Xm=X'-m;
        %X=Xm';
        [U,S,V] = svd(X,'econ');    %singular value decomposition
        S=diag(S);                      %get singular values
        allS(c,:)=S(1:numberofS);       %store them in matrix allS
    end
    allSs_final=[allSs_final allS];     %concatenate singular values
end
Xfinal=dataset(allSs_final);    % create PLS-Toolbox dataset
Xfinal.label{2}=samplenames;    % add sample names
Xfinal.class{1}=classes;        % add vector of classes
variablenames=[]    % create variable names
for i=1:size(included_segments,2)
    a=num2str(repmat(included_segments(i),numberofS,1))
    b=num2str(repmat('_',numberofS,1));
    c=num2str([1:numberofS]')
    d=[a b c]
    e=str2cell(d)
    variablenames=[variablenames;e]
end
Xfinal.label{2}=variablenames;
%% final PCA
% scaling (PLS-Toolbox function)
[Xfinal_scaled,sp] = preprocess('calibrate','classcentroidscale',Xfinal);
pca(Xfinal_scaled,10)       % pca (PLS-Toolbox function)
```

BIBLIOGRAPHY

BIBLIOGRAPHY

Abdi, H., Valentin, D., Chollet, S., and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food quality and preference*, 18(4):627–640.

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, 5(2):149–179.

Abrahamse, C. E. and Bartowsky, E. J. (2012). Timing of malolactic fermentation inoculation in Shiraz grape must and wine: influence on chemical composition. *World Journal of Microbiology and Biotechnology*, 28(1):255–265.

Adutwum, L. A. and Harynuk, J. J. (2014). Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis. *Analytical chemistry*, 86(15):7726–7733.

Aggio, R., Villas, S. G., and Ruggiero, K. (2011). Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics*, 27(16):2316–2318.

Allen, M. S., Lacey, M. J., Harris, R. L., and Brown, W. V. (1991). Contribution of methoxypyrazines to Sauvignon blanc wine aroma. *American Journal of Enology and Viticulture*, 42(2):109–112.

Alves, R., Nascimento, A., and Nogueira, J. (2005). Characterization of the aroma profile of Madeira wine by sorptive extraction techniques. *Analytica Chimica Acta*, 546(1):11–21.

Amigo, J. M., Popielarz, M. J., Callejón, R. M., Morales, M. L., Troncoso, A. M., Petersen, M. A., and Toldam-Andersen, T. B. (2010a). Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *Journal of Chromatography A*, 1217(26):4422–4429.

Amigo, J. M., Skov, T., and Bro, R. (2010b). ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics. *Chemical reviews*, 110(8):4582–4605.

Amigo, J. M., Skov, T., Bro, R., Coello, J., and Maspoch, S. (2008). Solving GC-MS problems with parafac2. *Trac Trends in Analytical Chemistry*, 27(8):714–725.

Andersson, C. A. and Bro, R. (2000). The N-way Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, 52(1):1–4.

Antalick, G. (2010). *Bilan biochimique et sensoriel des modifications de la note fruitée des vins rouges lors de la fermentation malolactique: rôle particulier des esters*. PhD thesis, Bordeaux 2.

Antalick, G., Perello, M., and de Revel, G. (2010a). Changes in wine secondary metabolite composition by the timing of inoculation with lactic acid bacteria: Impact on wine aroma. In *Proceedings of the 3rd International Symposium MACROWINE 2010 on Macromolecules and Secondary Metabolites in Grapevine and Wines*, pages 143–148. Universita di Torin Torino, Italy.

Antalick, G., Perello, M.-C., and de Revel, G. (2010b). Development, validation and application of a specific method for the quantitative determination of wine esters by headspace-solid-phase microextraction-gas chromatography–mass spectrometry. *Food chemistry*, 121(4):1236–1245.

Antalick, G., Perello, M.-C., and de Revel, G. (2012). Characterization of fruity aroma modifications in red wines during malolactic fermentation. *Journal of agricultural and food chemistry*, 60(50):12371–12383.

Antonelli, A., Castellari, L., Zambonelli, C., and Carnacini, A. (1999). Yeast influence on volatile composition of wines. *Journal of Agricultural and Food Chemistry*, 47(3):1139–1144.

Arapitsas, P., Della Corte, A., Gika, H., Narduzzi, L., Mattivi, F., and Theodoridis, G. (2016). Studying the effect of storage conditions on the metabolite content of red wine using HILIC LC-MS based metabolomics. *Food chemistry*, 197:1331–1340.

Arapitsas, P., Scholz, M., Vrhovsek, U., Di Blasi, S., Bartolini, A. B., Masuero, D., Perenzoni, D., Rigo, A., and Mattivi, F. (2012). A metabolomic approach to the study of wine micro-oxygenation. *PLoS One*, 7(5):e37783.

Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., and Mattivi, F. (2014). The influence of storage on the chemical age of red wines. *Metabolomics*, 10(5):816–832.

Arbulu, M., Sampedro, M., Gómez-Caballero, A., Goicolea, M., and Barrio, R. (2015). Untargeted metabolomic analysis using liquid chromatography quadrupole time-of-flight mass spectrometry for non-volatile profiling of wines. *Analytica chimica acta*, 858:32–41.

Atanassov, I., Hvarleva, T., Rusanov, K., Tsvetkov, I., and Atanassov, A. (2009). Wine metabolite profiling: possible application in winemaking and grapevine breading in bulgaria. *Biotechnology & Biotechnological Equipment*, 23(4):1449–1452.

Ballabio, D., Skov, T., Leardi, R., and Bro, R. (2008). Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques. *Journal of chemometrics*, 22(8):457–463.

Barcenas, P., Elortondo, F. P., and Albisu, M. (2004). Projective mapping in sensory analysis of ewes milk cheeses: A study on consumers and trained panel performance. *Food Research International*, 37(7):723–729.

Barclay, V., Bonner, R., and Hamilton, I. (1997). Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Analytical Chemistry*, 69(1):78–90.

Bartowsky, E. J. and Henschke, P. A. (2004). The buttery attribute of wine - diacetyl - desirability, spoilage and beyond. *International journal of food microbiology*, 96(3):235–252.

Behrends, V., Tredwell, G. D., and Bundy, J. G. (2011). A software complement to AMDIS for processing GC-MS metabolomic data. *Analytical biochemistry*, 415(2):206–208.

Bennetzen, J. and Hall, B. (1982). The primary structure of the Saccharomyces cerevisiae gene for alcohol dehydrogenase. *Journal of Biological Chemistry*, 257(6):3018–3025.

Bisson, L. F., Waterhouse, A. L., Ebeler, S. E., Walker, M. A., and Lapsley, J. T. (2002). The present and future of the international wine industry. *Nature*, 418(6898):696–699.

Black, C., Parker, M., Siebert, T., Capone, D., and Francis, I. (2015). Terpenoids and their role in wine flavour: recent advances. *Australian Journal of Grape and Wine Research*, 21(S1):582–600.

Boido, E., Medina, K., Farina, L., Carrau, F., Versini, G., and Dellacassa, E. (2009). The effect of bacterial strain and aging on the secondary volatile metabolites produced during malolactic fermentation of Tannat red wine. *Journal of agricultural and food chemistry*, 57(14):6271–6278.

Bonino, M., Schellino, R., Rizzi, C., Aigotti, R., Delfini, C., and Baiocchi, C. (2003). Aroma compounds of an Italian wine (Ruché) by HS–SPME analysis coupled with GC–ITMS. *Food Chemistry*, 80(1):125–133.

Borges, C. R. (2007). Concept for facilitating analyst-mediated interpretation of qualitative chromatographic-mass spectral data: an alternative to manual examination of extracted ion chromatograms. *Analytical chemistry*, 79(13):4805–4813.

Börner, J., Buchinger, S., and Schomburg, D. (2007). A high-throughput method for microbial metabolome analysis using gas chromatography/mass spectrometry. *Analytical biochemistry*, 367(2):143–151.

Boulton, R. (2001). The copigmentation of anthocyanins and its role in the color of red wine: a critical review. *American Journal of Enology and Viticulture*, 52(2):67–87.

Bouteille, R., Cordelle, S., Laval, C., Tournier, C., Lecanu, B., This, H., and Schlich, P. (2013). Sensory exploration of the freshness sensation in plain yoghurts and yoghurt-like products. *Food Quality and Preference*, 30(2):282–292.

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for experimenters: design, innovation, and discovery*, volume 2. Wiley-Interscience New York.

Brereton, R. G. (1995). Tutorial review. Deconvolution of mixtures by factor analysis. *Analyst*, 120(9):2313–2336.

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171.

Bro, R. (1998a). Least squares algorithms under unimodality and non-negativity constraints. *Journal of Chemometrics*, 12:223–247.

Bro, R. (1998b). *Multi-way analysis in the food industry: models, algorithms, and applications*. PhD thesis, Københavns Universitet.

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

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

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

Bruce, S. J., Jonsson, P., Antti, H., Cloarec, O., Trygg, J., Marklund, S. L., and Moritz, T. (2008). Evaluation of a protocol for metabolic profiling studies on human blood plasma by combined ultra-performance liquid chromatography/mass spectrometry: From extraction to data analysis. *Analytical Biochemistry*, 372(2):237–249.

Carroll, K. (1961). Quantitative estimation of peak areas in gas-liquid chromatography. *Nature*, 191:377–378.

Cartier, R., Rytz, A., Lecomte, A., Poblete, F., Krystlik, J., Belin, E., and Martin, N. (2006). Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map. *Food quality and preference*, 17(7):562–571.

Castillo, S., Gopalacharyulu, P., Yetukuri, L., and Orešič, M. (2011). Algorithms and tools for the preprocessing of LC–MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, 108(1):23–32.

Castro, C. C., Gunning, C., Oliveira, C. M., Couto, J. A., Teixeira, J. A., Martins, R. C., and Ferreira, A. C. S. (2012). Saccharomyces cerevisiae Oxidative Response Evaluation by Cyclic Voltammetry and Gas Chromatography–Mass Spectrometry. *Journal of agricultural and food chemistry*, 60(29):7252–7261.

Castro, C. C., Martins, R. C., Teixeira, J. A., and Ferreira, A. C. S. (2014). Application of a high-throughput process analytical technology metabolomics pipeline to Port wine forced ageing process. *Food chemistry*, 143:384–391.

Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., and Rodrick, G. E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science & Technology*, 20(11):557–566.

Chatonnet, P., Dubourdieu, D., and Boidron, J. (1995). The influence of Brettanomyces/Dekkera sp. yeasts and lactic acid bacteria on the ethylphenol content of red wines. *American Journal of Enology and Viticulture*, 46(4):463–468.

Chatonnet, P., Viala, C., and Dubourdieu, D. (1997). Influence of polyphenolic components of red wines on the microbial synthesis of volatile phenols. *American Journal of Enology and Viticulture*, 48(4):443–448.

Chessel, D. and Hanafi, M. (1996). Analyses de la co-inertie de K nuages de points. *Revue de Statistique Applique*, 44(2):35–60.

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

Christensen, J. H., Mortensen, J., Hansen, A. B., and Andersen, O. (2005b). Chromatographic preprocessing of GC–MS data for analysis of complex chemical mixtures. *Journal of Chromatography A*, 1062(1):113–123.

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

Clarke, R. J. and Bakker, J. (2004). *Wine flavour chemistry*. Wiley Online Library.

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

Conterno, L., Aprea, E., Franceschi, P., Viola, R., and Vrhovsek, U. (2013). Overview of Dekkera bruxellensis behaviour in an ethanol-rich environment using untargeted and targeted metabolomic approaches. *Food research international*, 51(2):670–678.

Cordella, C. B. and Bertrand, D. (2014). SAISIR: a new general chemometric toolbox. *TrAC Trends in Analytical Chemistry*, 54:75–82.

Costello, P., Francis, I., and Bartowsky, E. (2012). Variations in the effect of malolactic fermentation on the chemical and sensory properties of Cabernet Sauvignon wine: Interactive influences of Oenococcus oeni strain and wine matrix composition. *Australian Journal of Grape and Wine Research*, 18(3):287–301.

Croley, T. R., White, K. D., Callahan, J. H., and Musser, S. M. (2012). The chromatographic role in high resolution mass spectrometry for non-targeted analysis. *Journal of The American Society for Mass Spectrometry*, 23(9):1569–1578.

Cubero-Leon, E., Peñalver, R., and Maquet, A. (2014). Review on metabolomics for food authentication. *Food Research International*, 60:95–107.

Dairou, V. and Sieffermann, J.-M. (2002). A comparison of 14 jams characterized by conventional profile and a quick original method, the flash profile. *Journal of food science*, 67(2):826–834.

Danielsson, R., Bäckström, D., and Ullsten, S. (2006). Rapid multivariate analysis of LC/GC/CE data (single or multiple channel detection) without prior peak alignment. *Chemometrics and intelligent laboratory systems*, 84(1):33–39.

Daszykowski, M., Danielsson, R., and Walczak, B. (2008). No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis–mass spectrometry. *Journal of Chromatography A*, 1192(1):157–165.

Daszykowski, M. and Walczak, B. (2011). Methods for the exploratory analysis of two-dimensional chromatographic signals. *Talanta*, 83(4):1088–1097.

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

De la Calle García, D., Reichenbächer, M., Danzer, K., Hurlbeck, C., Bartzsch, C., and Feller, K.-H. (1997). Investigations on wine bouquet components by solid-phase microextration-capillary gas chromatography (SMPE-CGC) using different fibers. *Journal of High Resolution Chromatography*, 20(12):665–668.

De Mora, S., Eschenbruch, R., Knowles, S., and Spedding, D. (1986). The formation of dimethyl sulphide during fermentation using a wine yeast. *Food Microbiology*, 3(1):27–32.

De Vos, C., Tikunov, Y., Bovy, A., and Hall, R. (2008). Flavour metabolomics: Holistic versus targeted approaches in flavour research. In *Expression of Multidisciplinary Flavour Science. Proceedings of the 12th Weurman Symposium. Interlaken, Switzerland: Zürcher Hochschule für Angewandte and Institut Für Chemie und Biologische Chemie*, pages 573–580.

Dehlholm, C., Brockhoff, P. B., Meinert, L., Aaslyng, M. D., and Bredie, W. L. (2012). Rapid descriptive sensory methods–comparison of free multiple sorting,

partial napping, napping, flash profiling and conventional profiling. *Food Quality and Preference*, 26(2):267–277.

Delarue, J. and Sieffermann, J.-M. (2004). Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food quality and preference*, 15(4):383–392.

Denis, C. L., Ferguson, J., and Young, E. (1983). mRNA levels for the fermentative alcohol dehydrogenase of Saccharomyces cerevisiae decrease upon growth on a nonfermentable carbon source. *Journal of Biological Chemistry*, 258(2):1165–1171.

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

Dromey, R., Stefik, M. J., Rindfleisch, T. C., and Duffield, A. M. (1976). Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. *Analytical Chemistry*, 48(9):1368–1375.

Du, G., Zhan, J., Li, J., You, Y., Zhao, Y., and Huang, W. (2014). Effect of Grapevine Age on the Aroma Compounds in 'Beihong' Wine. *South African Journal of Enology and Viticulture*, 33(1):7–13.

Du Toit, M. and Pretorius, I. S. (2000). Microbial spoilage and preservation of wine: using weapons from natures own arsenal–a review. *South African Journal of Enology and Viticulture*, 21(Special Issue):74–96.

Dunn, W. B. and Ellis, D. I. (2005). Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4):285–294.

Durante, C., Bro, R., and Cocchi, M. (2011). A classification tool for N-way array based on SIMCA methodology. *Chemometrics and Intelligent Laboratory Systems*, 106(1):73–85.

Durante, C., Cocchi, M., Grandi, M., Marchetti, A., and Bro, R. (2006). Application of N-PLS to gas chromatographic and sensory data of traditional balsamic vinegars of Modena. *Chemometrics and Intelligent Laboratory Systems*, 83(1):54–65.

Eliasson, M., Rannar, S., and Trygg, J. (2011). From data processing to multivariate validation-essential steps in extracting interpretable information from metabolomics data. *Current pharmaceutical biotechnology*, 12(7):996–1004.

Elpa, D., Durán-Guerrero, E., Castro, R., Natera, R., and Barroso, C. G. (2014). Development of a new stir bar sorptive extraction method for the determination of medium-level volatile thiols in wine. *Journal of separation science*, 37(14):1867–1872.

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

Escofier, B. and Pages, J. (1994). Multiple factor analysis (AFMULT package). *Computational statistics & data analysis*, 18(1):121–140.

Escofier, B. and Pagès, J. (2008). *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation*. Dunod.

Escudero, A., Campo, E., Fariña, L., Cacho, J., and Ferreira, V. (2007). Analytical characterization of the aroma of five premium red wines. Insights into the role of odor families and the concept of fruitiness of wines. *Journal of Agricultural and Food Chemistry*, 55(11):4501–4510.

Fedrizzi, B., Carlin, S., Franceschi, P., Vrhovsek, U., Wehrens, R., Viola, R., and Mattivi, F. (2012). D-optimal design of an untargeted HS-SPME-GC-TOF metabolite profiling method. *Analyst*, 137(16):3725–3731.

Felinger, A. (1998). *Data analysis and signal processing in chromatography*, volume 21. Elsevier.

Fernandes, L., Relva, A., da Silva, M. G., and Freitas, A. C. (2003). Different multidimensional chromatographic approaches applied to the study of wine malolactic fermentation. *Journal of Chromatography A*, 995(1):161–169.

Ferreira, V., Rapp, A., Cacho, J. F., Hastrich, H., and Yavas, I. (1993). Fast and quantitative determination of wine flavor compounds using microextraction with Freon 113. *Journal of Agricultural and Food Chemistry*, 41(9):1413–1420.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2):155–171.

Fiehn, O. (2003). Metabolic networks of Cucurbita maxima phloem. *Phytochemistry*, 62(6):875–886.

Flamini, R., Vedova, A. D., Panighel, A., Perchiazzi, N., and Ongarato, S. (2005). Monitoring of the principal carbonyl compounds involved in malolactic fermentation of wine by solid-phase microextraction and positive ion chemical ionization GC/MS analysis. *Journal of mass spectrometry*, 40(12):1558–1564.

Fleet, G. H. (1993). *Wine microbiology and biotechnology*. CRC Press.

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

Francis, I. and Newton, J. (2005). Determining wine aroma from compositional data. *Australian Journal of Grape and Wine Research*, 11(2):114–126.

Furbo, S. and Christensen, J. H. (2012). Automated peak extraction and quantification in chromatography with multichannel detectors. *Analytical chemistry*, 84(5):2211–2218.

Gámbaro, A., Boido, E., Zlotejablko, A., Medina, K., Lloret, A., Dellacassa, E., and Carrau, F. (2001). Effect of malolactic fermentation on the aroma properties of Tannat wine. *Australian Journal of Grape and Wine Research*, 7(1):27–32.

Gammacurta, M., Marchand, S., Albertin, W., Moine, V., and de Revel, G. (2014). Impact of yeast strain on ester levels and fruity aroma persistence during aging of bordeaux red wines. *Journal of agricultural and food chemistry*, 62(23):5378–5389.

Gionfriddo, E., Souza-Silva, É. A., and Pawliszyn, J. (2015). Headspace versus Direct Immersion Solid Phase Microextraction in Complex Matrixes: Investigation of Analyte Behavior in Multicomponent Mixtures. *Analytical chemistry*, 87(16):8448–8456.

Groat, M. and Ough, C. (1978). Effects of insoluble solids added to clarified musts on fermentation rate, wine composition, and wine quality. *American Journal of Enology and Viticulture*, 29(2):112–119.

Halket, J. M., Przyborowska, A., Stein, S. E., Mallard, W. G., Down, S., and Chalmers, R. A. (1999). Deconvolution gas chromatography/mass spectrometry of urinary organic acids–potential for pattern recognition and automated identification of metabolic disorders. *Rapid communications in mass spectrometry*, 13(4):279–284.

Harshman, R. A. (1972). PARAFAC2: Mathematical and technical notes. *UCLA working papers in phonetics*, 22:30–44.

Harshman, R. A. and Lundy, M. E. (1994). PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72.

Hastings, C. A., Norton, S. M., and Roy, S. (2002). New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 16(5):462–467.

Hayasaka, Y., MacNamara, K., Baldock, G. A., Taylor, R. L., and Pollnitz, A. P. (2003). Application of stir bar sorptive extraction for wine analysis. *Analytical and bioanalytical chemistry*, 375(7):948–955.

Horning, E. and Horning, M.-G. (1971). Metabolic profiles: gas-phase methods for analysis of metabolites. *Clinical Chemistry*, 17(8):802–809.

Houtman, A. and Du Plessis, C. (1981). The effect of juice clarity and several conditions promoting yeast growth on fermentation rate, the production of aroma components and wine quality. *South African Journal of Enology and Viticulture*, 2(2):71.

Howell, K. S., Cozzolino, D., Bartowsky, E. J., Fleet, G. H., and Henschke, P. A. (2006). Metabolic profiling as a tool for revealing Saccharomyces interactions during wine fermentation. *FEMS Yeast Research*, 6(1):91–101.

Hübschmann, H.-J. (2008). *Handbook of GC/MS: fundamentals and applications*. John Wiley & Sons.

Jacobson, D., Monforte, A. R., and Ferreira, A. C. S. (2013). Untangling the Chemistry of Port Wine Aging with the Use of GC-FID, Multivariate Statistics, and Network Reconstruction. *Journal of agricultural and food chemistry*, 61(10):2513–2521.

Jellema, R. H., Krishnan, S., Hendriks, M. M., Muilwijk, B., and Vogels, J. T. (2010). Deconvolution using signal segmentation. *Chemometrics and Intelligent Laboratory Systems*, 104(1):132–139.

Johnsen, L. G., Amigo, J. M., Skov, T., and Bro, R. (2014). Automated resolution of overlapping peaks in chromatographic data. *Journal of Chemometrics*, 28(2):71–82.

Johnson, K. J. and Synovec, R. E. (2002). Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 60(1):225–237.

Jonsson, P., Bruce, S. J., Moritz, T., Trygg, J., Sjöström, M., Plumb, R., Granger, J., Maibaum, E., Nicholson, J. K., Holmes, E., et al. (2005). Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyst*, 130(5):701–707.

Kaminski, E., Stawicki, S., and Wasowicz, E. (1974). Volatile flavor compounds produced by molds of Aspergillus, Penicillium, and Fungi imperfecti. *Applied Microbiology*, 27(6):1001–1004.

Karagiannis, S. and Lanaridis, P. (2002). Insoluble grape material present in must affects the overall fermentation aroma of dry white wines made from three grape cultivars cultivated in Greece. *Journal of food science*, 67(1):369–374.

Katajamaa, M., Miettinen, J., and Orešič, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636.

Keim, H., de Revel, G., Marchand, S., and Bertrand, A. (2002). Method for determining nitrogenous heterocycle compounds in wine. *Journal of agricultural and food chemistry*, 50(21):5803–5807.

Kennedy, J. (2010). Evaluation of replicated projective mapping of granola bars. *Journal of Sensory Studies*, 25(5):672–684.

Kiers, H. A. (1991). Hierarchical relations among three-way methods. *Psychometrika*, 56(3):449–470.

Knoll, C., Fritsch, S., Schnell, S., Grossmann, M., Krieger-Weber, S., du Toit, M., and Rauhut, D. (2012). Impact of different malolactic fermentation inoculation scenarios on Riesling wine aroma. *World Journal of Microbiology and Biotechnology*, 28(3):1143–1153.

Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C., and Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics*, 7(3):307–328.

Koek, M. M., Muilwijk, B., van der Werf, M. J., and Hankemeier, T. (2006). Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical chemistry*, 78(4):1272–1281.

Lacey, M. J., Allen, M. S., Harris, R. L., and Brown, W. V. (1991). Methoxypyrazines in Sauvignon blanc grapes and wines. *American Journal of Enology and Viticulture*, 42(2):103–108.

Laghi, L., Versari, A., Marcolini, E., and Parpinello, G. P. (2014). Metabonomic investigation by 1H-NMR to discriminate between red wines from organic and biodynamic grapes. *Food and Nutrition Sciences*, 2014.

Lambrechts, M. and Pretorius, I. (2000). Yeast and its importance to wine aroma - a review. *South African Journal of Enology and Viticulture*.

Lange, E., Gröpl, C., Schulz-Trieglaff, O., Leinenbach, A., Huber, C., and Reinert, K. (2007). A geometric approach for the alignment of liquid chromatography - mass spectrometry data. *Bioinformatics*, 23(13):i273–i281.

Lawless, H. T., Sheng, N., and Knoops, S. S. (1995). Multidimensional scaling of sorting data applied to cheese perception. *Food Quality and Preference*, 6(2):91–98.

Lay, D. (2002). *Linear Algebra and Its Applications*. Addison Wesley, 3 edition.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, 25(1):1–18.

Lee, J.-E., Hwang, G.-S., Lee, C.-H., and Hong, Y.-S. (2009). Metabolomics reveals alterations in both primary and secondary metabolites by wine bacteria. *Journal of agricultural and food chemistry*, 57(22):10772–10783.

Liang, H.-Y., Chen, J.-Y., Reeves, M., and Han, B.-Z. (2013). Aromatic and sensorial profiles of young Cabernet Sauvignon wines fermented by different Chinese autochthonous Saccharomyces cerevisiae strains. *Food research international*, 51(2):855–865.

Lilly, M., Lambrechts, M., and Pretorius, I. (2000). Effect of increased yeast alcohol acetyltransferase activity on flavor profiles of wine and distillates. *Applied and environmental microbiology*, 66(2):744–753.

Lommen, A. (2009). MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical chemistry*, 81(8):3079–3086.

Lommen, A. and Kools, H. J. (2012). MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics*, 8(4):719–726.

Lopez, R., Aznar, M., Cacho, J., and Ferreira, V. (2002). Determination of minor and trace volatile compounds in wine by solid-phase extraction and gas chromatography with mass spectrometric detection. *Journal of Chromatography A*, 966(1):167–177.

Lorho, G., Westad, F., and Bro, R. (2006). Generalized correlation loadings: extending correlation loadings to congruence and to multi-way models. *Chemometrics and intelligent laboratory systems*, 84(1):119–125.

Losada, M. M., Andrés, J., Cacho, J., Revilla, E., and López, J. F. (2011). Influence of some prefermentative treatments on aroma composition and sensory evaluation of white Godello wines. *Food chemistry*, 125(3):884–891.

Loscos, N., Hernandez-Orte, P., Cacho, J., and Ferreira, V. (2007). Release and formation of varietal aroma compounds during alcoholic fermentation from nonfloral grape odorless flavor precursors fractions. *Journal of Agricultural and Food Chemistry*, 55(16):6674–6684.

Louw, C., La Grange, D., Pretorius, I., and Van Rensburg, P. (2006). The effect of polysaccharide-degrading wine yeast transformants on the efficiency of wine processing and wine flavour. *Journal of biotechnology*, 125(4):447–461.

Lopez-Rituerto, E., Savorani, F., Avenoza, A., Busto, J. H., Peregrina, J. M., and Engelsen, S. B. (2012). Investigations of La Rioja terroir for wine production using 1H NMR metabolomics. *Journal of agricultural and food chemistry*, 60(13):3452–3461.

Luedemann, A., Strassburg, K., Erban, A., and Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography - mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737.

Lukić, I., Banović, M., Peršurić, D., Radeka, S., and Sladonja, B. (2006). Determination of volatile compounds in grape distillates by solid-phase extraction and gas chromatography. *Journal of Chromatography A*, 1101(1):238–244.

Lytra, G., Tempere, S., Le Floch, A., de Revel, G., and Barbe, J.-C. (2013). Study of sensory interactions among red wine fruity esters in a model solution. *Journal of agricultural and food chemistry*, 61(36):8504–8513.

Lytra, G., Tempere, S., Revel, G. d., and Barbe, J.-C. (2012). Impact of perceptive interactions on red wine fruity aroma. *Journal of agricultural and food chemistry*, 60(50):12260–12269.

Maeder, M. and Zilian, A. (1988). Evolving factor analysis, a new multivariate technique in chromatography. *Chemometrics and Intelligent Laboratory Systems*, 3(3):205–213.

Mallard, W. G. (2014). AMDIS in the chemical weapons convention. *Analytical and bioanalytical chemistry*, 406(21):5075–5086.

Mallouchos, A., Komaitis, M., Koutinas, A., and Kanellaki, M. (2003). Wine fermentations by immobilized and free cells at different temperatures. Effect of immobilization and temperature on volatile by-products. *Food Chemistry*, 80(1):109–113.

Marais, J. (1983). Terpenes in the aroma of grapes and wines: a review. *South African Journal of Enology and Viticulture*, 4(2):49–60.

Marais, J. and Pool, H. (1980). Effect of storage time and temperature on the volatile composition and quality of dry white table wines. *Vitis*, 19(2):151–164.

Marchand, S., Almy, J., and de Revel, G. (2011). The Cysteine Reaction with Diacetyl under Wine-Like Conditions: Proposed Mechanisms for Mixed Origins of 2-Methylthiazole, 2-Methyl-3-thiazoline, 2-Methylthiazolidine, and 2,4,5-Trimethyloxazole. *Journal of food science*, 76(6):C861–C868.

Marchand, S., de Revel, G., and Bertrand, A. (2000). Approaches to wine aroma: release of aroma compounds from reactions between cysteine and carbonyl compounds in wine. *Journal of agricultural and food chemistry*, 48(10):4890–4895.

Marchand, S., de Revel, G., Vercauteren, J., and Bertrand, A. (2002). Possible mechanism for involvement of cysteine in aroma production in wine. *Journal of agricultural and food chemistry*, 50(21):6160–6164.

Martens, H. and Martens, M. (2001). *Multivariate analysis of quality. An introduction*. IOP Publishing.

Mazerolles, G., Hanafi, M., Dufour, E., Bertrand, D., and Qannari, E. (2006). Common components and specific weights analysis: a chemometric method for dealing with complexity of food products. *Chemometrics and Intelligent Laboratory Systems*, 81(1):41–49.

Mcdaniel, M., Henderson, L. A., Watson, B. T., and Heatherbell, D. (1987). Sensory panel training and screening for descriptive analysis of the aroma of Pinot Noir wine fermented by several strains of malolactic bacteria. *Journal of Sensory Studies*, 2(3):149–167.

Mendes-Pinto, M. M. (2009). Carotenoid breakdown products the - norisoprenoids - in wine aroma. *Archives of Biochemistry and Biophysics*, 483(2):236–245.

Meyer, M. R., Peters, F. T., and Maurer, H. H. (2010). Automated mass spectral deconvolution and identification system for GC-MS screening for drugs, poisons, and metabolites in urine. *Clinical Chemistry*, 56(4):575–584.

Mohler, R. E., Dombek, K. M., Hoggard, J. C., Pierce, K. M., Young, E. T., and Synovec, R. E. (2007). Comprehensive analysis of yeast metabolite GC×GC–TOFMS data: combining discovery-mode and deconvolution chemometric software. *Analyst*, 132(8):756–767.

Moio, L., Ugliano, M., Gambuti, A., Genovese, A., and Piombino, P. (2004). Influence of clarification treatment on concentrations of selected free varietal aroma compounds and glycoconjugates in Falanghina (*Vitis vinifera L.*) must and wine. *American journal of enology and viticulture*, 55(1):7–12.

Monforte, A. R., Jacobson, D., and Silva Ferreira, A. (2015). Chemiomics: Network Reconstruction and Kinetics of Port Wine Aging. *Journal of agricultural and food chemistry*, 63(9):2576–2581.

Muller, C. J., Kepner, R. E., and Webb, A. D. (1973). Lactones in wines - a review. *American Journal of Enology and Viticulture*, 24(1):5–9.

Muñoz, D., Peinado, R. A., Medina, M., and Moreno, J. (2007). Biological aging of sherry wines under periodic and controlled microaerations with Saccharomyces cerevisiae var. capensis: Effect on odorant series. *Food Chemistry*, 100(3):1188–1195.

Murphy, K. R., Wenig, P., Parcsi, G., Skov, T., and Stuetz, R. M. (2012). Characterizing odorous emissions using new software for identifying peaks in chemometric models of gas chromatography–mass spectrometry datasets. *Chemometrics and Intelligent Laboratory Systems*, 118:41–50.

Nestrud, M. A. and Lawless, H. T. (2008). Perceptual mapping of citrus juices using projective mapping and profiling data from culinary professionals and consumers. *Food quality and preference*, 19(4):431–438.

Nestrud, M. A. and Lawless, H. T. (2010). Perceptual mapping of apples and cheeses using projective mapping and sorting. *Journal of Sensory Studies*, 25(3):390–405.

Nicolini, G., Moser, S., Román, T., Mazzi, E., and Larcher, R. (2015). Effect of juice turbidity on fermentative volatile compounds in white wines. *Vitis*, 50(3):131.

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

Niu, W., Knight, E., Xia, Q., and McGarvey, B. D. (2014). Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments. *Journal of Chromatography A*, 1374:199–206.

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

Pagès, J. (2003). Recueil direct de distances sensorielles: application à l'évaluation de dix vins blancs du Val-de-Loire. *Sciences des aliments*, 23(5-6):679–688.

Pagès, J. (2005a). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food quality and preference*, 16(7):642–649.

Pagès, J. (2005b). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food quality and preference*, 16(7):642–649.

Pagès, J., Cadoret, M., and Lê, S. (2010). The sorted napping: A new holistic approach in sensory evaluation. *Journal of Sensory Studies*, 25(5):637–658.

Palomo, E. S., Hidalgo, M. D.-M., Gonzalez-Vinas, M., and Pérez-Coello, M. (2005). Aroma enhancement in wines from different grape varieties using exogenous glycosidases. *Food chemistry*, 92(4):627–635.

Patel, S. and Shibamoto, T. (2002). Effect of different strains of *Saccharomyces cerevisiae* on production of volatiles in Napa Gamay wine and Petite Sirah wine. *Journal of agricultural and food chemistry*, 50(20):5649–5653.

Perestrelo, R., Fernandes, A., Albuquerque, F., Marques, J., and Câmara, J. (2006). Analytical characterization of the aroma of Tinta Negra Mole red wine: Identification of the main odorants compounds. *Analytica Chimica Acta*, 563(1):154–164.

Perrin, L. and Pagès, J. (2009). Construction of a product space from the ustra-flash profiling method: application to 10 red wines from the loire valley. *Journal of Sensory Studies*, 24(3):372–395.

Perrin, L., Symoneaux, R., Maître, I., Asselin, C., Jourjon, F., and Pagès, J. (2008). Comparison of three sensory methods for use with the Napping® procedure: Case of ten wines from Loire valley. *Food Quality and Preference*, 19(1):1–11.

Peyrot des Gachons, C., Tominaga, T., and Dubourdieu, D. (2000). Measuring the aromatic potential of Vitis vinifera L. Cv. Sauvignon blanc grapes by assaying S-cysteine conjugates, precursors of the volatile thiols responsible for their varietal aroma. *Journal of agricultural and food chemistry*, 48(8):3387–3391.

Picard, M., Thibon, C., Redon, P., Darriet, P., de Revel, G., and Marchand, S. (2015). Involvement of dimethyl sulfide and several polyfunctional thiols in the aromatic expression of the aging bouquet of red Bordeaux wines. *Journal of Agricultural and Food Chemistry*, 63(40):8879–8889.

Pierce, K. M., Hoggard, J. C., Hope, J. L., Rainey, P. M., Hoofnagle, A. N., Jack, R. M., Wright, B. W., and Synovec, R. E. (2006). Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts. *Analytical Chemistry*, 78(14):5068–5075.

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

Piñeiro, Z., Natera, R., Castro, R., Palma, M., Puertas, B., and Barroso, C. (2006). Characterisation of volatile fraction of monovarietal wines: Influence of winemaking practices. *Analytica chimica acta*, 563(1):165–172.

Pinu, F. R., Edwards, P. J., Jouanneau, S., Kilmartin, P. A., Gardner, R. C., and Villas-Boas, S. G. (2014). Sauvignon blanc metabolomics: grape juice metabolites affecting the development of varietal thiols and other aroma compounds in wines. *Metabolomics*, 10(4):556–573.

Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11(1):395.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rankine, B. and Pococx, K. (1969). Influence of yeast starin, grape variety and other factors; and taste thresholds. *Vitis*, 8:23–37.

Rapp, A. and Mandery, H. (1986). Wine aroma. *Experientia*, 42(8):873–884.

Reynolds, A. G., Pearson, E. G., De Savigny, C., Coventry, J., and Strommer, J. (2008). Interactions of vine age and reflective mulch upon berry, must, and wine composition of five *Vitis vinifera* cultivars. *International Journal of Fruit Science*, 7(4):85–119.

Ribéreau-Gayon, P., Dubourdieu, D., Donèche, B., and Lonvaud, A. (2006). *Handbook of Enology, The microbiology of wine and vinifications*, volume 1. John Wiley & Sons.

Ribéreau-Gayon, P., Glories, Y., Maujean, A., and Dubourdieu, D. (2000). Handbook of Enology. The chemistry of wine and stabilisation and treatments.

Risvik, E., McEwan, J. A., Colwill, J. S., Rogers, R., and Lyon, D. H. (1994). Projective mapping: A tool for sensory analysis and consumer research. *Food quality and preference*, 5(4):263–269.

Risvik, E., McEwan, J. A., and Rødbotten, M. (1997). Evaluation of sensory profiling and projective mapping data. *Food Quality and Preference*, 8(1):63–71.

Riu-Aumatell, M., Bosch-Fusté, J., López-Tamames, E., and Buxaderas, S. (2006). Development of volatile compounds of cava (Spanish sparkling wine) during long ageing time in contact with lees. *Food Chemistry*, 95(2):237–242.

Robinson, A. L., Boss, P. K., Heymann, H., Solomon, P. S., and Trengove, R. D. (2011a). Development of a sensitive non-targeted method for characterizing the wine volatile profile using headspace solid-phase microextraction comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *Journal of Chromatography A*, 1218(3):504–517.

Robinson, A. L., Boss, P. K., Heymann, H., Solomon, P. S., and Trengove, R. D. (2011b). Influence of yeast strain, canopy management, and site on the volatile composition and sensory attributes of Cabernet Sauvignon wines from Western Australia. *Journal of agricultural and food chemistry*, 59(7):3273–3284.

Rocha, S., Ramalheira, V., Barros, A., Delgadillo, I., and Coimbra, M. A. (2001). Headspace solid phase microextraction (SPME) analysis of flavor compounds in wines. Effect of the matrix volatile composition in the relative response factors in a wine model. *Journal of Agricultural and Food Chemistry*, 49(11):5142–5151.

Rochfort, S., Ezernieks, V., Bastian, S. E., and Downey, M. O. (2010). Sensory attributes of wine influenced by variety and berry shading discriminated by NMR metabolomics. *Food Chemistry*, 121(4):1296–1304.

Rodrigues, J. A., Barros, A. S., Carvalho, B., Brandão, T., Gil, A. M., and Ferreira, A. C. S. (2011). Evaluation of beer deterioration by gas chromatography–mass spectrometry/multivariate analysis: A rapid tool for assessing beer composition. *Journal of chromatography A*, 1218(7):990–996.

Rodríguez, M. C., Sánchez, G. H., Sobrero, M. S., Schenone, A. V., and Marsili, N. R. (2013). Determination of mycotoxins (aflatoxins and ochratoxin A) using fluorescence emission-excitation matrices and multivariate calibration. *Microchemical Journal*, 110:480–484.

Roland, A., Schneider, R., Razungles, A., and Cavelier, F. (2011). Varietal thiols in wine: discovery, analysis and applications. *Chemical reviews*, 111(11):7355–7376.

Romano, P., Fiore, C., Paraggio, M., Caruso, M., and Capece, A. (2003). Function of yeast species and strains in wine flavour. *International journal of food microbiology*, 86(1):169–180.

Ross, C. F., Weller, K. M., and Alldredge, J. R. (2012). Impact of serving temperature on sensory properties of red wine as evaluated using projective mapping by a trained panel. *Journal of Sensory Studies*, 27(6):463–470.

Roullier-Gall, C., Witting, M., Tziotis, D., Ruf, A., Gougeon, R., and Schmitt-Kopplin, P. (2015). Integrating analytical resolutions in non-targeted wine metabolomics. *Tetrahedron*, 71(20):2983–2990.

Salkind, N. J. (2006). *Encyclopedia of measurement and statistics*. Sage Publications.

Sánchez-Palomo, E., Diaz-Maroto, M. C., and Perez-Coello, M. S. (2005). Rapid determination of volatile compounds in grapes by HS-SPME coupled with GC–MS. *Talanta*, 66(5):1152–1157.

Sandra, P., Tienpont, B., Vercammen, J., Tredoux, A., Sandra, T., and David, F. (2001). Stir bar sorptive extraction applied to the determination of dicarboximide fungicides in wine. *Journal of chromatography A*, 928(1):117–126.

Santos, B., Pollonio, M., Cruz, A., Messias, V., Monteiro, R., Oliveira, T., Faria, J., Freitas, M., and Bolini, H. (2013). Ultra-flash profile and projective mapping for describing sensory attributes of prebiotic mortadellas. *Food Research International*, 54(2):1705–1711.

Sauvageot, F. and Vivier, P. (1997). Effects of malolactic fermentation on sensory properties of four Burgundy wines. *American Journal of Enology and Viticulture*, 48(2):187–192.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Schiffman, S. S., Reynolds, M. L., Young, F. W., and Carroll, J. D. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. Academic press New York.

Schmidtke, L. M., Blackman, J. W., Clark, A. C., and Grant-Preece, P. (2013). Wine metabolomics: objective measures of sensory properties of semillon from GC-MS profiles. *Journal of agricultural and food chemistry*, 61(49):11957–11967.

Schneider, R., Charrier, F., Razungles, A., and Baumes, R. (2006). Evidence for an alternative biogenetic pathway leading to 3-mercaptohexanol and 4-mercapto-4-methylpentan-2-one in wines. *Analytica Chimica Acta*, 563(1):58–64.

Sefton, M., Skouroumounis, G. K., Massywestropp, R. A., and Williams, P. (1989). Norisoprenoids in Vitis vinifera white wine grapes and the identification of a precursor of damascenone in these fruits. *Australian Journal of Chemistry*, 42(12):2071–2084.

Selli, S., Canbas, A., Cabaroglu, T., Erten, H., and Günata, Z. (2006). Aroma components of cv. Muscat of Bornova wines and influence of skin contact treatment. *Food Chemistry*, 94(3):319–326.

Silva Ferreira, A. C., Monforte, A. R., Teixeira, C. S., Martins, R., Fairbairn, S., and Bauer, F. F. (2014). Monitoring Alcoholic Fermentation: An Untargeted Approach. *Journal of agricultural and food chemistry*, 62(28):6784–6793.

Silva Ferreira, A. C., Rodrigues, P., Hogg, T., and Guedes de Pinho, P. (2003). Influence of some technological parameters on the formation of dimethyl sulfide, 2-mercaptoethanol, methionol, and dimethyl sulfone in port wines. *Journal of Agricultural and Food Chemistry*, 51(3):727–732.

Singleton, V., Sieberhagen, H., De Wet, P., and Van Wyk, C. (1975). Composition and sensory qualities of wines prepared from white grapes by fermentation with and without grape solids. *American Journal of Enology and Viticulture*, 26(2):62–69.

Sinkov, N. A. and Harynuk, J. J. (2011). Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta*, 83(4):1079–1087.

Sinkov, N. A. and Harynuk, J. J. (2013). Three-dimensional cluster resolution for guiding automatic chemometric model optimization. *Talanta*, 103:252–259.

Sinkov, N. A., Johnston, B. M., Sandercock, P. M. L., and Harynuk, J. J. (2011). Automated optimization and construction of chemometric models based on highly variable raw chromatographic data. *Analytica chimica acta*, 697(1):8–15.

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

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

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, 78(3):779–787.

Souza-Silva, É. A., Jiang, R., Rodríguez-Lafuente, A., Gionfriddo, E., and Pawliszyn, J. (2015). A critical review of the state of the art of solid-phase microextraction of complex matrices I. Environmental analysis. *TrAC Trends in Analytical Chemistry*, 71:224–235.

Sparkman, O. D., Penton, Z., and Kitson, F. G. (2011). *Gas Chromatography and Mass Spectrometry: A Practical Guide*. Academic Press.

Springer, A., Riedl, J., Esslinger, S., Roth, T., Glomb, M., and Fauhl-Hassek, C. (2014). Validated modeling for German white wine varietal authentication based

on headspace solid-phase microextraction online coupled with gas chromatography mass spectrometry fingerprinting. *Journal of agricultural and food chemistry*, 62(28):6844–6851.

Stanimirova, I., Walczak, B., Massart, D., Simeonov, V., Saby, C., and Di Crescenzo, E. (2004). STATIS, a three-way method for data analysis. Application to environmental data. *Chemometrics and Intelligent Laboratory Systems*, 73(2):219–233.

Stein, S., Mirokhin, Y., Tchekhovskoi, D., and Mallard, G. (2008). *NIST Mass Spectral Search Program*. National Institute of Standards and Technology, Gaithersburg, MD.

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

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R. C. (2008). Sensory Evaluation by Quantitative Descriptive Analysis. In Gacula, M. C., editor, *Descriptive Sensory Analysis in Practice*, chapter 1.3, pages 23–34. Food & Nutrition Press, Inc.

Šuklje, K., Antalick, G., Coetzee, Z., Schmidtke, L., Baša Česnik, H., Brandt, J., Toit, W., Lisjak, K., and Deloire, A. (2014). Effect of leaf removal and ultraviolet radiation on the composition and sensory perception of Vitis vinifera L. cv. Sauvignon Blanc wine. *Australian Journal of Grape and Wine Research*, 20(2):223–233.

Sumby, K. M., Grbin, P. R., and Jiranek, V. (2010). Microbial modulation of aromatic esters in wine: current knowledge and future prospects. *Food Chemistry*, 121(1):1–16.

Swiegers, J., Bartowsky, E., Henschke, P., and Pretorius, I. (2005). Yeast and bacterial modulation of wine aroma and flavour. *Australian Journal of grape and wine research*, 11(2):139–173.

Swiegers, J. H., Kievit, R. L., Siebert, T., Lattey, K. A., Bramley, B. R., Francis, I. L., King, E. S., and Pretorius, I. S. (2009). The influence of yeast on the aroma of Sauvignon Blanc wine. *Food Microbiology*, 26(2):204–211.

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

Tapp, H. S. and Kemsley, E. K. (2009). Notes on the practical utility of OPLS. *TrAC Trends in Analytical Chemistry*, 28(11):1322–1327.

Tarr, P. T., Dreyer, M. L., Athanas, M., Shahgholi, M., Saarloos, K., and Second, T. P. (2013). A metabolomics based approach for understanding the influence of terroir in Vitis Vinifera L. *Metabolomics*, 9(1):170–177.

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

Teillet, E., Schlich, P., Urbano, C., Cordelle, S., and Guichard, E. (2010). Sensory methodologies and the taste of water. *Food Quality and Preference*, 21(8):967–976.

Teofilo, R. F., Martins, J. P. A., and Ferreira, M. (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics*, 23(1):32–48.

Theodoridis, G. A., Gika, H. G., Want, E. J., and Wilson, I. D. (2012). Liquid chromatography–mass spectrometry based global metabolite profiling: a review. *Analytica chimica acta*, 711:7–16.

Thuillier, B., Valentin, D., Marchal, R., and Dacremont, C. (2015). Pivot© profile: A new descriptive method based on free description. *Food Quality and Preference*, 42:66–77.

Timberlake, C. and Bridle, P. (1976). Interactions between anthocyanins, phenolic compounds, and acetaldehyde and their significance in red wines. *American Journal of Enology and Viticulture*, 27(3):97–105.

Toffali, K., Zamboni, A., Anesi, A., Stocchero, M., Pezzotti, M., Levi, M., and Guzzo, F. (2011). Novel aspects of grape berry ripening and post-harvest withering revealed by untargeted LC-ESI-MS metabolomics analysis. *Metabolomics*, 7(3):424–436.

Tomasi, G., Savorani, F., and Engelsen, S. B. (2011). icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A*, 1218(43):7832–7840.

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

Tominaga, T., Peyrot des Gachons, C., and Dubourdieu, D. (1998). A New Type of Flavor Precursors in Vitis v inifera L. cv. Sauvignon Blanc: S-Cysteine Conjugates. *Journal of Agricultural and Food Chemistry*, 46(12):5215–5219.

Torri, L., Dinnella, C., Recchia, A., Naes, T., Tuorila, H., and Monteleone, E. (2013). Projective mapping for interpreting wine aroma differences as perceived by naïve and experienced assessors. *Food Quality and Preference*, 29(1):6–15.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics*, 16(3):119–128.

Ugliano, M. and Moio, L. (2005). Changes in the concentration of yeast-derived volatile compounds of red wine during malolactic fermentation with four commercial starter cultures of Oenococcus oeni. *Journal of agricultural and food chemistry*, 53(26):10134–10139.

Valentin, D., Chollet, S., Lelievre, M., and Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology*, 47(8):1563–1578.

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

Van Nederkassel, A., Daszykowski, M., Eilers, P., and Vander Heyden, Y. (2006). A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210.

Varela, P. and Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International*, 48(2):893–908.

Vestner, J., Malherbe, S., Du Toit, M., Nieuwoudt, H. H., Mostafa, A., Górecki, T., Tredoux, A. G., and De Villiers, A. (2011). Investigation of the volatile composition of pinotage wines fermented with different malolactic starter cultures using comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC-TOF-MS). *Journal of agricultural and food chemistry*, 59(24):12732–12744.

Vivó-Truyols, G., Torres-Lapasió, J., Van Nederkassel, A., Vander Heyden, Y., and Massart, D. (2005). Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection. *Journal of Chromatography A*, 1096(1):133–145.

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

Want, E. and Masson, P. (2011). Processing and Analysis of GC/LC-MS-Based Metabolomics Data. In Metz, T. O., editor, *Metabolic Profiling*, volume 708 of *Methods in Molecular Biology*, pages 277–298. Humana Press.

Wehrens, R., Weingart, G., and Mattivi, F. (2014). metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *Journal of Chromatography B*, 966:109–116.

Weldegergis, B. T. and Crouch, A. M. (2008). Analysis of volatiles in Pinotage wines by stir bar sorptive extraction and chemometric profiling. *Journal of agricultural and food chemistry*, 56(21):10225–10236.

Williams, A. A. and Langron, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, 35(5):558–568.

Williams, J. T., Ough, C., and Berg, H. (1978). White wine composition and quality as influenced by method of must clarification. *American Journal of Enology and Viticulture*, 29(2):92–96.

Winterhalter, P. (1991). 1,1,6-Trimethyl-1,2-dihydronaphthalene (TDN) formation in wine. 1. Studies on the hydrolysis of 2,6,10,10-tetramethyl-1-oxaspiro [4.5] dec-6-ene-2,8-diol rationalizing the origin of TDN and related C13 norisoprenoids in Riesling wine. *Journal of agricultural and Food Chemistry*, 39(10):1825–1829.

Winterhalter, P., Sefton, M., and Williams, P. (1990). Volatile C13-norisoprenoid compounds in Riesling wine are generated from multiple precursors. *American journal of enology and viticulture*, 41(4):277–283.

Wishart, D. S. (2008). Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, 19(9):482–493.

Wold, H. et al. (1966). Estimation of principal components and related models by iterative least squares. In Krishnaiah, P. R., editor, *Multivariate analysis : proceedings of an International symposium held in Dayton, Ohio, June 14-19, 1965*. New York; London: Academic.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern recognition*, 8(3):127–139.

Wold, S., Johansson, E., Jellum, E., Bjørnson, I., and Nesbakken, R. (1981). Application of SIMCA multivariate data analysis to the classification of gas chromatographic profiles of human brain tissues. *Analytica Chimica Acta*, 133(3):251–259.

Yu, P. and Pickering, G. J. (2008). Ethanol difference thresholds in wine and the influence of mode of evaluation and wine style. *American journal of enology and viticulture*, 59(2):146–152.

Zalacain, A., Alonso, G., Lorenzo, C., Iniguez, M., and Salinas, M. (2004). Stir bar sorptive extraction for the analysis of wine cork taint. *Journal of Chromatography A*, 1033(1):173–178.

Zalacain, A., Marín, J., Alonso, G., and Salinas, M. (2007). Analysis of wine primary aroma compounds by stir bar sorptive extraction. *Talanta*, 71(4):1610–1615.

Zerzucha, P., Kazura, M., de Beer, D., Joubert, E., Schulze, A. E., Beelders, T., de Villiers, A., and Walczak, B. (2013). A new concept for variance analysis of hyphenated chromatographic data avoiding signal warping. *Journal of Chromatography A*, 1291:64–72.

Zufferey, V. and Maigre, D. (2008). Age de la vigne II. Influence sur la qualité des raisins et des vins. *Revue suisse de viticulture, arboriculture, horticulture*, 40(4):241–245.