



UNIVERSITAT<sub>DE</sub>  
BARCELONA

# Data processing for Life Sciences measurements with hyphenated Gas Chromatography-Ion Mobility Spectrometry

Sergio Oller Moreno



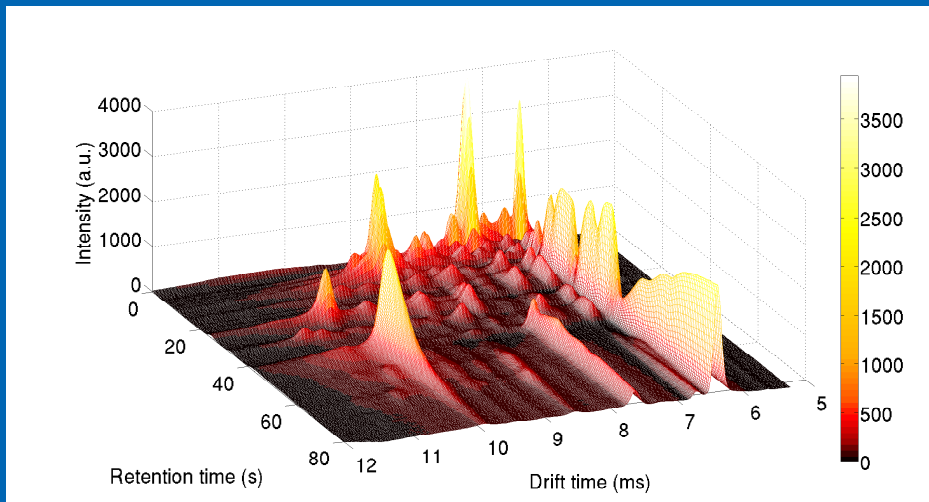
Aquesta tesi doctoral està subjecta a la llicència [Reconeixement- Compartigual 3.0. Espanya de Creative Commons](#).

Esta tesis doctoral está sujeta a la licencia [Reconocimiento - Compartirlqual 3.0. España de Creative Commons](#).

This doctoral thesis is licensed under the [Creative Commons Attribution-ShareAlike 3.0. Spain License](#).

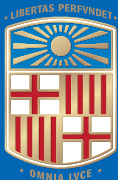
Tesi doctoral

# Data processing for Life Sciences measurements with hyphenated Gas Chromatography-Ion Mobility Spectrometry



**Sergio Oller Moreno**

Supervisor: Antonio Pardo Martínez  
Dept. Enginyeria Electrònica i Biomèdica



UNIVERSITAT DE  
BARCELONA



This page is intentionally left blank



FACULTAT DE FÍSICA


Departament d'Enginyeria Electrònica i Biomèdica

MEMÒRIA PRESENTADA PER OPTAR AL GRAU DE DOCTOR PER LA  
UNIVERSITAT DE BARCELONA

Programa de Doctorat en Enginyeria i Ciències Aplicades  
(RD 99/2011)

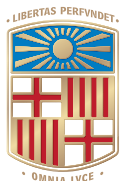
DATA PROCESSING FOR LIFE SCIENCES  
MEASUREMENTS WITH HYPHENATED GAS  
CHROMATOGRAPHY-ION MOBILITY SPECTROMETRY

Sergio Oller Moreno

 [orcid.org/0000-0002-8994-1549](https://orcid.org/0000-0002-8994-1549)

Director i tutor: Dr. Antonio Pardo Martínez

Barcelona, 2017



UNIVERSITAT DE  
BARCELONA



# Agraïments

Ha estat una tesi llarga. I no hauria estat una tesi sense l'ajuda, el suport i la paciència de moltes persones, al llarg de tots aquests anys. Sé que no les menciono a totes (faria una tesi d'agraïments), i segur que en falten de molt importants, però us agrairé la vostra benevolència i que no m'ho tingueu en compte.

Ana, gracias por tu infinita paciencia, por tus ánimos, por tu soporte y por tu alegría. Gracias por motivarme a seguir, por tener siempre una sonrisa. Por los consejos, las reflexiones y los ánimos. Gracias por querer formar equipo conmigo, nos espera una vida de aventuras y de días increíbles. Muero por disfrutarlos a tu lado.

A mis padres, no sé si os lo digo lo suficiente: Gracias por hacerme la persona que soy, gracias por todos los momentos en los que me habéis apoyado, animado y educado. Gracias por vuestra paciencia. Me siento muy afortunado de teneros. Gracias por todo lo que me queréis.

Gracias también a toda mi enorme familia. En especial a mi Abuela Inma, a la que admiro profundamente por su increíble capacidad de adaptación y por su habilidad para escribir ¡quién la tuviera! A mis tíos y primos por ambos lados, Oller y Moreno, y también a Morgen (thanks for your constant support!) y a todos los Rasmussen-Mitjana. He intentado explicaros (a todos) lo que hacía, aunque al principio ni siquiera yo lo sabía, espero que con esta tesis quede un poco más claro... Un agradecimiento muy cariñoso también a mi nueva familia política en México con la que me siento como en casa: Gracias por vuestros corazones, que sé que desde Durango nos animáis a no desfallecer y a terminar los doctorados. Tengo ganas de veros pronto.

En el entorno más profesional, quiero dar las gracias al Dr. Antonio Pardo, Toni, por dirigir esta tesis. Por tu ayuda al principio de la tesis cuando estando en el laboratorio venías y te sentabas a mi lado a discutir los problemas que tenía, por las mini reuniones semanales que teníamos, por ayudarme a estructurar este



trabajo. Y sobretodo gracias por tus ánimos constantes, tus “cap problema”, tus “tienes que sacarte esto ya”, tus “vamos a discutirlo con Santi”... sin ese apoyo yo no estaría escribiendo esto ahora. Gracias por haber confiado en mí.

Gracias también al Dr. Santiago Marco, Santi, por todas esas charlas, discusiones, ideas y propuestas. Por haber guiado y por haber apretado, por haber buscado colaboraciones y por haberlas encontrado. “Todo puede mejorarse y siempre se puede hacer más, pero lo mejor es enemigo de lo bueno así que hay que acabar ya”. Más allá de todas las mejoras que puedan hacerse, gracias por el reconocimiento del trabajo realizado a lo largo de estos años. Todos los estudiantes de doctorado del grupo te preocupan, y sé que te has preocupado mucho por mí. De ahora en adelante espero darte solo alegrías (¡y publicaciones!).

Gracias a todos los demás miembros del grupo, presentes y pasados. Agustín, Javier Adrián, Guillermo, Dani, Didier, Sergi “output” Udina, Erola (quins grans consells amb el Luís al bar!), Pablo, Quiroga, Dani (con cara de Rafa), Víctor, Milad, Alex, Rudys, Núria, Ariadna, Nil, Raquel, Jordi, Raquel, Yia (I want to visit China!), Raquel... y también aquellos con los que he coincidido en estancias más breves: Selda, Patricio, Marimar, Begum, Saeedeh, it was a pleasure meeting you. Mucha gente ha pasado por el grupo, y los que seguramente me habré olvidado... pero tienen un lugar especial en mi corazón Ana Guamán, Marta, Javi, Sílvia, Luís, Sara, Juanma, Ana (aquí también, como compañera de trabajo) y Raquel... Gracias por todos los consejos, debates, discusiones, pero sobretodo por todas las risas, todos los momentos, los chistes, los cafés, los bikinis, todas las cervezas y las patatas bravas que jamás podré olvidar. Sois geniales.

Quiero agradecer también el esfuerzo del personal de administración, divulgación y soporte a la investigación del IBEC: os agradezco habernos divertido organizando actividades escolares, las sobras de algunos pica-pica, la paciencia y diligencia de Fran y de Juli en las incidencias técnicas, y la sonrisa y la amabilidad que me llevo cada vez que paso por administración. Es un placer ir a veros.

Al Departament d'Electrònica i Biomèdica vull agrair les bones tradicions com les calçotades i els picatrònics, i fer passar un xic de vergonya a qui es presenti. En especial vull agrair al Manolo i a la Nuria amb qui al principi del doctorat acostumàvem a fer uns bons esmorzars al bar, els riures i xerrades que fèiem.

Y como no agradecer la colaboración con la Dra. Lourdes Arce de la Universidad de Córdoba, con Rocío, con Nati y con todo el grupo. Gracias por compartir vuestra experiencia y conocimiento en química, en instrumentación y en IMS, y gracias por acogerme tan amablemente.

There is all the nice people as well I met at NIHS, that gave me a lot of positive feelings and from whom I learnt a lot. I would like to thank all of you, but you are

all too many, so please forgive me if your name is missing... Thanks Ornella, Ivan, Mapy, Martine, Sofia, Loïc, Antonio, Federico, Filippo, Polina, Ondine, François, Sebastiano, Seu-Ping, Eugenia, Martin, Jim, Filomena, Oksana, Tobias, India, Fiona, Laeticia, Irina, Hélène, Tony, Jerome, Armand, Michaël, Yann, Anita, Alessio, Laurence, Nicolas, Delphine, Olivier, Bernhard, Richard, Radovan, Soren, Julie, Jérôme, Sofia, Sarah, Rolf, Lynne, Laura, Danielle, Orélie, Jörg, Ed, Eric, and I am deeply sorry for the ones I am forgetting. I should have said “everyone”, but that was not specific enough to express my gratitude. I had a great time and I am looking forward to seeing you soon again.

Thanks as well to all the rstats community, who is building a great set of tools for doing data analysis. Brilliant people who devote some of their time to open source packages so everyone can benefit and work in a more comfortable way: Yihui Xie, Kirill Müller, Jenny Bryan, Hadley Wickham, Jim Hester, Kara Woo, and so many others! Thanks for taking the time of replying to issues, reviewing pull requests and providing always constructive feedback. I hope I haven't stolen too much of your time.

Oh, and a special “*spasiba*” to Aleksandra Elbakian. Without her contributions and probably without the contributions of so many other volunteers I would not have been able to read all the things I have read.

I evidentment, tornant a un entorn més personal, també agrair a tots els amics que han estat al costat al llarg d'aquests anys. Al Miquel i al Jordi, amb qui he conviscut i compartit el dia a dia, les alegries i les penes, les viciades, les partides al “PhD the Game” que fan que el doctorat sigui fàcil d'explicar... i a l'Elia, la Sílvia, la Paula i el Ferré, que ens anem veient a ratxes, però riem molt cada cop que ho fem. I al Miquel, que havent tornat ell de Finlàndia i estant els dos lliures de doctorats, espero que puguem quedar més sovint :-).

A la colla d'amics Físics pel món, amb qui he pogut compartir carrera i/o màster, penes de doctorat, calçotajes, sopars de nadal, birres, quinieles,... i que cada cop estem més dispersos pel planeta. Fa nostàlgia veure que les distàncies mica en mica creixen, però alhora és una bonica mostra que tots anem creixent. Un pingüí m'ha portat tot just avui una sèrie de fotografies carregades de records de part vostra. M'heu emocionat. Gràcies.

No puc acabar aquest agraïment sense mencionar la Beth i el Ferran, sense ells no hauria triat física i probablement no seria on sóc ara. L'impacte que mestres i professors de secundària i batxillerat teniu en les vides de la gent que passa per les vostres mans és meravellós, amb un punt esfereïdor si un hi pensa massa. Gràcies per haver educat a tantes i tantes persones cada any.

Gràcies a tots.



# Abstract

Recent progress in analytical chemistry instrumentation has increased the amount of data available for analysis. This progress has been encompassed by computational improvements, that have enabled new possibilities to analyze larger amounts of data. These two factors have allowed to analyze more complex samples in multiple life science fields, such as biology, medicine, pharmacology, or food science.

One of the techniques that has benefited from these improvements is Gas Chromatography - Ion Mobility Spectrometry (GC-IMS). This technique is useful for the detection of Volatile Organic Compounds (VOCs) in complex samples. Ion Mobility Spectrometry is an analytical technique for characterizing chemical substances based on the velocity of gas-phase ions in an electric field. It is able to detect trace levels of volatile chemicals reaching for some analytes ppb concentrations. While the instrument has moderate selectivity it is very fast in the analysis, as an ion mobility spectrum can be acquired in tenths of milliseconds. As it operates at ambient pressure, it is found not only as laboratory instrumentation but also in-site, to perform screening applications. For instance it is often used in airports for the detection of drugs and explosives. To enhance the selectivity of the IMS, especially for the analysis of complex samples, a gas chromatograph can be used for sample pre-separation at the expense of the length of the analysis.

While there is better instrumentation and more computational power, better algorithms are still needed to exploit and extract all the information present in the samples. In particular, GC-IMS has not received much attention compared to other analytical techniques. In this work we address some of the data analysis issues for GC-IMS: With respect to the pre-processing, we explore several baseline estimation methods and we suggest a variation of Asymmetric Least Squares, a popular baseline estimation technique, that is able to cope with signals that present large peaks or large dynamic range. This baseline estimation method is used in Gas Chromatography - Mass Spectrometry signals as well, as it suits both

techniques. Furthermore, we also characterize spectral misalignments in a several months long study, and propose an alignment method based on monotonic cubic splines for its correction. Based on the misalignment characterization we propose an optimal time span between consecutive calibrant samples.

We then explore the usage of Multivariate Curve Resolution methods for the deconvolution of overlapped peaks and their extraction into pure components. We propose the use of a sliding window in the retention time axis to extract the pure components from smaller windows. The pure components are tracked through the windows. This approach is able to extract analytes with lower response with respect to MCR, compounds that have a low variance in the overall matrix

Finally we apply some of these developments to real world applications, on a dataset for the prevention of fraud and quality control in the classification of olive oils, measured with GC-IMS, and on data for biomarker discovery of prostate cancer by analyzing the headspace of urine samples with a GC-MS instrument.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Life Sciences samples . . . . .	3
1.2.1	The Human Volatolome . . . . .	3
1.2.2	Food quality control and fraud prevention . . . . .	4
1.3	Instrumentation . . . . .	5
1.3.1	Ion Mobility Spectrometry . . . . .	5
1.3.1.1	Operating Principle . . . . .	5
1.3.1.2	IMS signal characteristics . . . . .	8
1.3.1.3	Other IMS variants . . . . .	11
1.3.2	Gas chromatography . . . . .	11
1.3.2.1	Gas chromatography – Mass Spectrometry . . . . .	14
1.3.2.2	Gas chromatography – Ion Mobility Spectrometry . . . . .	15
1.4	Data analysis for IMS and hyphenated instrumentation . . . . .	18
1.4.1	Preprocessing . . . . .	18
1.4.1.1	Denoising methods . . . . .	19
1.4.1.2	Baseline estimation methods . . . . .	19
1.4.1.3	Alignment methods . . . . .	20
1.4.1.4	Normalization . . . . .	22
1.4.2	Feature extraction: Peak deconvolution . . . . .	23
1.4.3	Outlier detection and lab quality control . . . . .	24
1.4.4	Classification and Regression . . . . .	25
1.4.5	Validation . . . . .	26
1.5	Objectives . . . . .	27
<b>2</b>	<b>Preprocessing</b>	<b>29</b>
2.1	Denoising . . . . .	30
2.2	Baseline estimation . . . . .	31
2.2.1	Data description . . . . .	32

2.2.1.1	Synthetic dataset . . . . .	32
2.2.1.2	Real samples . . . . .	35
2.2.2	Baseline estimation method description . . . . .	36
2.2.2.1	Original Asymmetric Least Squares . . . . .	36
2.2.2.2	airPLS correction . . . . .	38
2.2.2.3	Proposed method: psalsa . . . . .	39
2.2.3	Results . . . . .	39
2.2.3.1	Synthetic chromatograms . . . . .	39
2.2.3.2	Real samples . . . . .	42
2.2.4	Discussion and remarks . . . . .	42
2.3	Alignment . . . . .	44
2.3.1	Dataset . . . . .	45
2.3.2	Methodology . . . . .	47
2.3.3	Results and Discussion . . . . .	49
2.3.3.1	Drift time correction . . . . .	50
2.3.3.2	Retention time correction . . . . .	52
<b>3</b>	<b>Sliding Window Multivariate Curve Resolution for GC-IMS data</b>	<b>57</b>
3.1	Introduction . . . . .	58
3.1.1	Blind Source Separation techniques . . . . .	58
3.1.2	Multivariate Curve Resolution Alternating Least Squares . . . . .	59
3.1.3	Proposed technique: Sliding Window Multivariate-Curve Resolution . . . . .	61
3.2	Materials and Methods . . . . .	62
3.2.1	Description of the samples . . . . .	62
3.2.2	Analytical methods . . . . .	63
3.2.3	Pre-processing . . . . .	63
3.2.4	Sliding Window MCR . . . . .	64
3.3	Results and Discussion . . . . .	65
<b>4</b>	<b>Applications</b>	<b>73</b>
4.1	Olive oil quality analysis with GC-IMS . . . . .	73
4.1.1	Experimental Protocol . . . . .	74
4.1.2	Data analysis methodology . . . . .	75
4.1.3	Results . . . . .	77
4.1.3.1	Preprocessing . . . . .	77
4.1.3.2	Model training and accuracy . . . . .	78
4.2	Prostate cancer biomarker discovery . . . . .	82
4.2.1	Methods . . . . .	82
4.2.2	Data analysis . . . . .	83
<b>5</b>	<b>Conclusions</b>	<b>89</b>

<b>Appendix</b>	<b>90</b>
<b>A Resum de la tesi</b>	<b>91</b>
A.1 Introducció . . . . .	91
A.1.1 Instrumentació . . . . .	93
A.2 Preprocessat . . . . .	95
A.2.1 Correcció de la línia de base . . . . .	96
A.2.2 Alineat . . . . .	98
A.3 Resolució multivariant de corbes en finestra mòbil (SW-MCR) . . .	101
A.4 Aplicacions . . . . .	103
A.4.1 Anàlisi de qualitat d'oli d'oliva mitjançant GC-IMS . . . .	103
A.4.2 Cerca de biomarcadors de càncer de pròstata en els volàtils de la orina . . . . .	105
<b>B Conclusions</b>	<b>109</b>
<b>C Publications</b>	<b>111</b>
C.1 Contributions to open source packages . . . . .	111
C.2 Publications . . . . .	111
C.3 Oral Presentations in conferences . . . . .	112
C.4 Posters . . . . .	112
<b>Bibliography</b>	<b>129</b>





# List of Tables

1.1	Number of Volatile Organic Compounds per body fluid . . . . .	3
2.1	Time distribution of 44 calibrant samples . . . . .	46
2.2	Raw peak positions in GC-IMS calibrant dataset . . . . .	49
2.3	Peak positions after drift time correction . . . . .	51
2.4	Comparison of the estimated retention times . . . . .	54
3.1	Comparison of 22 peak positions by MCR-ALS and SW-MCR deconvolution . . . . .	67
A.1	Nombre de compostos volàtils per fluid corporal . . . . .	92
A.2	Posicions originals dels pics en els calibrants de GC-IMS . . . . .	99
A.3	Peak positions after drift time correction . . . . .	99
A.4	Comparativa dels temps de retenció estimats . . . . .	100
A.5	Comparativa de 22 posicions de pics fent servir deconvolucions MCR-ALS i SW-MCR . . . . .	104



# List of Figures

1.1	Ion Mobility Spectrometer diagram . . . . .	5
1.2	IMS spectra measured with a <sup>63</sup> Nickel radioactive source of a blank and a mixture of ethanol and acetone. . . . .	9
1.3	IMS RIP/monomer/dimer peak intensities example . . . . .	10
1.4	Gas Chromatograph schema . . . . .	12
1.5	Total Ion Chromatogram of a human urine sample . . . . .	13
1.6	GC-MS of the headspace of a human urine sample. . . . .	14
1.7	Gas Chromatography – Ion Mobility Spectrometer diagram . . . . .	16
1.8	Region of a MCC-IMS olive oil sample . . . . .	17
1.9	Reverse Reactant Ion Peak example . . . . .	17
2.1	Region of GC spectra with manual peak boundaries marked . . . . .	31
2.2	Synthetic chromatogram . . . . .	33
2.3	Generalized exponential peaks . . . . .	34
2.4	Real urine samples . . . . .	35
2.5	Asymmetric Least Squares baseline fit in successive iterations . . . . .	38
2.6	Region of a synthetic sample showing different baseline estimations. . . . .	40
2.7	Comparison of the three methods for synthetic chromatograms . . . . .	41
2.8	Performance of the psalsa baseline estimation algorithm for several exponents . . . . .	41
2.9	Comparison of the baseline corrections applied to real samples . . . . .	42
2.10	Detailed region of the comparison of the baseline corrections methods on real samples . . . . .	43
2.11	Data analysis flow chart for the alignment study . . . . .	47
2.12	Raw peak positions before alignment . . . . .	49
2.13	Peak positions after drift time correction . . . . .	50
2.14	Correction factor for the drift time alignment . . . . .	52
2.15	Warping functions for linear and cubic retention time corrections . . . . .	53
2.16	Reverse RIP alignment result across multiple samples. . . . .	55

2.17	Alignment error depending on age of the previous calibrant . . . . .	56
3.1	Spectra tracking diagram with SW-MCR . . . . .	62
3.2	Region of a MCC-IMS olive oil sample . . . . .	65
3.3	Reverse Reactant Ion Peak in MCC-IMS olive oil sample . . . . .	66
3.4	MCC-IMS sample region (contour plot) . . . . .	68
3.5	Pure spectra and concentration profiles resolved by MCR-ALS . . . . .	69
3.6	Pure spectra and concentration profiles resolved by SW-MCR . . . . .	70
3.7	Tracked compounds along several windows . . . . .	70
3.8	Extracted Reactant Ion Peak using SW-MCR . . . . .	71
3.9	Spurious compounds rejection . . . . .	72
4.1	Data analysis flow chart for olive oil discrimination . . . . .	75
4.2	Double cross-validation diagram for the olive oil analysis . . . . .	76
4.3	Baseline correction for the olive oil application . . . . .	78
4.4	RIP alignment correction for the olive oil application . . . . .	79
4.5	Retention time alignment for olive oil samples . . . . .	79
4.6	Classification rate vs model complexity . . . . .	80
4.7	Score projection for the first two latent variables . . . . .	81
4.8	Loadings for the first latent variable . . . . .	81
4.9	Baseline subtraction effect . . . . .	84
4.10	Peak detection in GC-MS samples . . . . .	85
4.11	Angle between reference and mass spectra . . . . .	86
4.12	Boxplot of detected peaks . . . . .	86
A.1	Diagrama d'un cromatògraf de gasos acoblat amb un IMS . . . . .	93
A.2	Regió d'una mostra MCC-IMS d'oli d'oliva . . . . .	95
A.3	Comparació de tres mètodes per cromatogrames sintètics . . . . .	98
A.4	Comparació de l'estimació de la línia de base en mostres reals . . . . .	99
A.5	Diagrama de seguiment d'espectres amb SW-MCR . . . . .	103
A.6	Diagrama de doble validació creuada . . . . .	105
A.7	Projecció d'scores i loadings . . . . .	106
A.8	Detecció de pics a mostres GC-MS. . . . .	107
A.9	Boxplot de pics . . . . .	107

# Chapter 1

## Introduction

### 1.1 Motivation

The aim of this thesis is the development of algorithms for the analysis of complex gas-phase samples using hyphenated instrumentation, in particular Gas Chromatography – Ion Mobility Spectrometry (GC-IMS).

The development of algorithms to extract information from analytical chemistry instrumentation has existed since the 1970s, under the name of chemometrics. Chemometrics data typically has been obtained from UV/visible spectroscopy, chromatography, mass spectrometry, nuclear magnetic resonance, and atomic emission/absorption experiments, using multivariate data analysis techniques. While IMS and GC-IMS have received some attention, recent reviews (Hauschild et al., 2012) highlight the need for better algorithms and tailored data analysis methods. In the last decade, the rise of \*omics fields has partially absorbed the chemometrics area (Geladi and Hopke, 2008) as indeed most of the multivariate data analysis methods traditionally used by chemometricians are now also used in \*omics applications.

This thesis will explore the existing techniques for data analysis both in GC-IMS instrumentation and in other similar analytical techniques, and propose enhancements, modifications and adaptations of those algorithms to suit better the needs of GC-IMS. Currently, most statistical analysis of GC-IMS data rely on the use of either closed source software provided by the instrument manufacturer or a third party vendor (e.g. VisualNow<sup>1</sup>), or the use of techniques scattered

---

<sup>1</sup><http://www.bs-analytik.de/en/products/software-vocan-visualnow.htm>

among several packages and articles. The algorithms developed in this thesis will be published under a modular open source toolbox that hopefully can be freely extended and reused by the community.

Ion Mobility Spectrometry as an analytical tool has been useful in several fields: On the **security** industry, IMS is being used on a daily basis in airports for the detection of drugs and explosives (Eiceman et al., 2013). On the **pharmaceutical** industry, IMS is used in cleaning validations, to ensure that there is no carryover effect or cleaning agents left between consecutive batches of a pharmaceutical product (O'Donnell et al., 2008). It is used as well for the detection of triacetone triperoxide (TATP) in air in the industry (Räsänen et al., 2008). It is more and more being used as a research tool for the analysis of life sciences samples, in particular for the analysis of Volatile Organic Compounds (VOC).

The analysis of VOCs from life sciences samples is having increasing relevance in **medical** and **food quality** applications. Recent studies on VOCs, not only with GC-IMS but also with other analytical instrumentation, show their potential for the diagnostic of medical conditions and for food quality control.

For instance, medically, several research groups have shown that breath samples contain biomarkers for multiple conditions, such as: *pulmonary tuberculosis* (Phillips et al., 2010), *breast cancer* (Phillips et al., 2006), *lung cancer* (Buszewski et al., 2012, Fuchs et al. (2010)) or *chronic obstructive pulmonary disease* (Westhoff et al., 2010). Urine volatiles have been reported to possibly contribute to the diagnosis of *prostate cancer* (Cornu et al., 2011), (Khalid et al., 2015) and volatiles in vaginal discharge fluids are being used for diagnosing *vaginosis* (Karpas et al., 2012a).

In **food quality control** applications, wine volatiles have shown potential to be used for the detection of “tainted wine” (Karpas et al., 2012b, Márquez-Sillero et al. (2011)), beer volatiles can be used for fermentation control (Vautz et al., 2006a) and olive oil volatiles can help in the prevention of olive oil quality fraud (Garrido-Delgado et al., 2012), among other applications (Karpas, 2013).

The analysis of these complex samples relies on the recent advances in hyphenated instrumentation (Sarker and Nahar, 2012). While typically ion mobility spectrometry is not adequate for the analysis of complex samples due to its lack of selectivity, a pre-separation using gas chromatography overcomes that limitation at the expense of both portability and speed of analysis. Additionally, the pre-separation information can be used as well for analyte identification. The larger amount of data generated by the hyphenated instrument requires a set of specific algorithms to extract all the sample information.

In order to provide a solid ground base for the description of the proposed algo-

rithms, in this introduction we will briefly describe typical life sciences samples of interest for GC-IMS applications. We will continue exploring the analytical instrumentation, focusing on the main instruments of interest (IMS and GC-IMS) but also with some details in Gas Chromatography – Mass Spectrometry, as some of the algorithms developed are also suitable and have been tested with data obtained from those instruments. Finally, we will explore the state of the art in the data analysis workflows for GC-IMS spectra and describe the goals of this work.

## 1.2 Life Sciences samples

### 1.2.1 The Human Volatolome

The analysis of VOCs in human fluids and the understanding of their role in the metabolic pathways is an important step for the early diagnosis of many medical conditions. Recent reports (de Lacy Costello et al., 2014) aim to provide a compendium of VOCs in human body fluids, giving an initial description of what is called the *human volatolome*. A summary of the number of VOC in the volatolome per type of fluid is given in table 1.1. The complexity of the samples, with hundreds of compounds per body fluid is clear.

Table 1.1: Number of VOCs per body fluid, as reported in (de Lacy Costello et al., 2014). The compendium is built from samples belonging to healthy subjects. It should be noted that this is not by far the total number of existing VOCs, as for instance the high number of VOCs reported for skin secretions is due to the larger number of existing sample preparation methods, and the low number of urine VOCs (compared to e.g. faeces) is reported to be related to the low concentrations of those VOCs in urine, not to the fact that they are missing.

Body fluid	Number of VOCs
Breath	872
Saliva	359
Blood	154
Milk	256
Skin secretions	532
Urine	279
Faeces	381

The complexity and variability of these samples is not only due to the high num-



ber of VOCs. For instance, if we focus on human breath samples, the VOCs come from the alveolar breath, that participates in the gas exchange process with the blood, in concentrations of the range of ppm or even ppb. In this case, the number of VOCs ranges in the high hundreds, but only a small number of them are common to everyone (Mukhopadhyay, 2004). Some reasons for this variability are that some of these VOCs are exogenous and come from the environment, for instance from pollution, trees or cleaning products. Other VOCs are endogenous and related to the physiology of the subject, and may highlight a medical condition of some sort. This complexity and variability is not specific of breath samples, on the contrary it is known to be common to all body fluids.

The analysis of the volatolome is far from complete, and there still are several concerns and open problems that need to be addressed if we aim to improve the translation of the research results to clinics. The major concern for this translation to succeed is the low reproducibility in pre-medical studies (Begley and Ioannidis, 2015). Therefore we need reliable data analysis tools that follow best practices for the discovery of biomarkers. There are calls in the scientific community that aim to remind and encourage the commitment to those best practices (Broadhurst and Kell, 2007) and this thesis aims to follow that direction, by providing reliable data analysis tools. Other concerns are the improvement of sample acquisition protocols, VOC extraction methods and instrumentation, that can provide new ways to for instance normalize samples, reducing subject variability.

In this thesis we will see the application of data processing algorithms to a GC-MS dataset of urine volatiles, with the aim of discovering prostate cancer biomarkers.

## 1.2.2 Food quality control and fraud prevention

The food industry is another field that benefits from the analysis of VOCs. Having fast and reliable methods for the discovery of food spoilage issues, to control food and beverage production or to assess the quality of raw ingredients and materials is desirable not only because of the economical savings, but also and more importantly because it can prevent food waste. Typically, headspace analyses require enrichment of the sample air and further lab analysis using spectroscopic methods (Conte et al., 1999). The most common standard analytical techniques (atomic absorption spectroscopy, gas chromatography, mass spectrometry) require sample preparation, are time consuming, and expensive (Vautz et al., 2006c). In the last decade there has been an increase of alternative methods to address or complement by screening methods some of these issues. As it has already been mentioned, there are reports of the use of Ion Mobility Spectrometry in food

freshness and spoilage (Raatikainen et al., 2005), process control in the beer industry (Vautz et al., 2006b), wine quality (Karpas et al., 2012b) and olive oil fraud prevention (Garrido-Delgado et al., 2012).

In this thesis we will see how the developed algorithms are applied to a GC-IMS dataset of olive oils, with the aim of classifying them according to their quality.

## 1.3 Instrumentation

### 1.3.1 Ion Mobility Spectrometry

Ion Mobility Spectrometry (IMS) is an analytical technique for characterizing chemical substances based on the velocity of gas-phase ions in an electric field (Eiceman et al., 2013). IMS technology is able to detect trace levels of volatile chemicals, reaching for some analytes ppb concentrations. The analysis is fast (one spectrum is acquired in tenths of milliseconds) and the instrument has moderate selectivity.

IMS started gaining popularity in the 1970s, coined under the “Plasma Chromatography” term. It was first used for explosives and chemical warfare agents detection in military applications (Karasek and Denney, 1974), (Ewing, 2001). However, over the past twenty years, IMS fields of application have widened and it is currently being used in a wide range of applications (Armenta et al., 2011) such as environmental (Karpas et al., 1991), (Baumbach et al., 1993); industrial (Budde, 1995); biomedical studies (Westhoff et al., 2010), (Baumbach, 2009); drug detection (Eatherton et al., 1986), (Nanji et al., 1987); security applications (Cline and Hobbs, 1972), food quality (Vautz et al., 2006c), (Garrido-Delgado et al., 2012); cosmetics (Zamora et al., 2011); and fraud detection (Alonso et al., 2008). The use of IMS for explosive and drugs detection has now a large market, as nowadays many international airports are using the technology on a day-to-day basis.

#### 1.3.1.1 Operating Principle

The ion mobility spectrometer is divided in the ionization region, the shutter grid, and the drift region. As shown in figure 1.1, the sample enters into the ionization region where an ionization source is responsible for ionizing the molecules in the sample at ambient pressure. The ionized samples enter the drift region when a shutter grid opens and travel through the drift tube, that typically consists of a stack made of metallic rings. These rings are set to a decreasing range of electrical

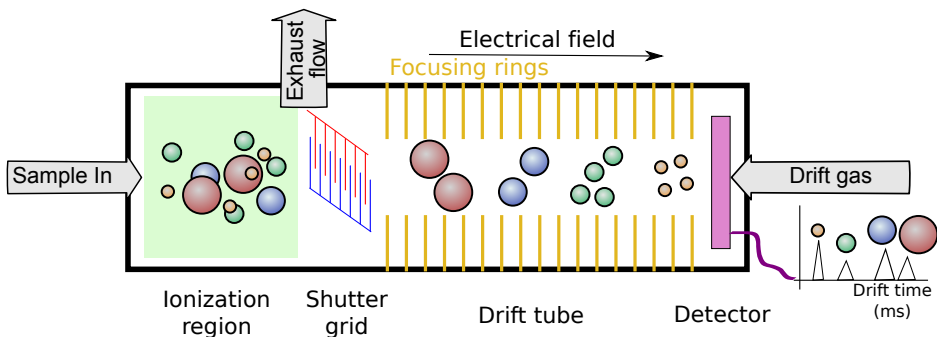


Figure 1.1: Ion Mobility Spectrometer diagram

potential creating a constant electrical field. The ions will be accelerated by this electrical field and their acceleration will depend on their mass and charge. The ions will be colliding with the neutral molecules of the drift gas (usually air or  $N_2$ ) that flows in the opposite direction. The number of collisions will depend on the cross section of the ions and the gas, with typical mean free path distances between collisions of  $10^{-7}$  m (Eiceman et al., 2013). The different ions will be separated in the drift tube by their electrical mobility, that depends on their mass, shape and charge among other factors, reaching a detector at the end of the drift tube (typically made of a Faraday plate). The time it takes for an ion to travel through the drift tube is the drift time of the ion. If the ratio of the electrical field with respect to the drift gas density is small  $E/N$ , the high number of collisions with the drift gas dissipate the energy acquired from the electrical field. This leads to a linear relation between the mean speed of the ion in the drift tube and the applied electrical field  $v_d = KE$ . This proportion  $K$  is known as electrical mobility and mainly depends on the reduced mass of the drift gas and the ion, the ion charge, the cross section, and the temperature.

To make sure the drift time is measured properly, ions are kept at the shutter grid before entering the drift tube. This grid is opened with a pulse of a short period of time, of the order of  $100 - 500 \mu\text{s}$ . Shorter pulses let pass less ions on an IMS scan reducing the sensitivity of the IMS. Longer pulses let pass more ions leading to more ion-ion repulsions that broaden the peaks. Knowing how the shape of the peaks is affected by the shutter grid is important in order to diagnose and understand instrumental issues.

While the drift time the ions take to travel through the drift tube  $t_d$  is the magnitude measured, the electrical mobility  $K$  is much more practical to work with, as it accounts for both the electrical field applied  $E$  and the length of the

drift tube  $L$  as shown in (1.1).

$$K = \frac{v_d}{E} = \frac{L}{Et_d} \quad (1.1)$$

Even more practical than the electrical mobility is the reduced electrical mobility, that also considers the pressure and temperature conditions so ease the comparison between experiments. The expression for the reduced mobility is given at (1.2), with the temperature expressed in Kelvin and the pressure in Torr (Eiceman et al., 2013).

$$K_0 = K \frac{273}{T} \frac{P}{760} \quad (1.2)$$

Still, even if the reduced mobility compensates for the electrical field and the tube length, and normalizes temperature and pressure variations, it does not account for all changes that can happen due to influences of temperature, pressure and gas composition on the ion identities and cross sections (Berant et al., 1989).

Depending on the IMS instrument, the pressure and temperature information may or may not be directly available. It is important to measure them if they are not available, as they can be used to control spectral misalignments as we will see later in the introduction 1.4.1.3.

### **Ionization sources**

Even if the separation of the sample components happens in the drift tube, the ionization source has a fundamental role in the IMS, as the choice of the ionization source can determine what ions can be formed. The most common for IMS are radioactive, corona discharge, photodischarge lamps and laser ionization sources (Eiceman et al., 2013).

Radioactive sources (typically beta-emitting  $^{63}\text{Ni}$ , beta-emitting Tritium or alpha-emitting Americium) are the most common source types, as they have advantages in terms of portability (no need of additional power sources and low maintenance). The beta radiation from Nickel of 67 keV ionizes the supporting atmosphere of the sample producing reactant ions. These ions generate more secondary electrons, and this process happens until generated ions lack the energy to ionize the supporting atmosphere. Compared to  $^{63}\text{Ni}$  sources, Tritium sources have less energy (18 keV) being less hazardous, while Americium is preferable when smaller volumes are required due to their short effective range. The main drawbacks of the use of radioactive sources are their environmental impact

the associated regulations on the use of radioactive instrumentation, that require specific permissions and controls.

Corona discharge sources are typically built with a sharp needle or wire separated few mm from a metal plate. A voltage difference of few kV is established between the plate and needle and when the discharge happens ions are formed in the gap between them. The formed ions are similar to those of a radioactive source and they contribute in further reactions with the sample. The main advantages of corona discharge sources are its simple design, no radioactivity and high ion currents. Its main disadvantages are the need of a high voltage power supply and a high maintenance due to the corrosion and erosion of the components. Specifically the corrosion of the needle degrades the stability of the source.

Both the radioactive and the corona discharge ionization sources use forms of indirect ionization, where the supporting atmosphere is ionized first and subsequent reactions ionize the sample. Photoionization lamps and lasers, on the contrary, use direct ionization. A photoionization lamp will emit photons from the excitation of its internal gas. These photons will impact with the sample molecules and, if the energy/frequency of the photon is right it will form cations of the sample molecules by removing the valence electron. Further reactions in the sample may happen but the reaction mechanisms for the ionization are not yet fully understood. The main advantage of photoionization is that by choosing the right internal gas in the lamp one can select the energy of the emitted photons ensuring some selectivity of the instrument. The main disadvantages are the need of an external power supply and the need of replacing the lamps periodically, as their lifespan is limited. Their ionization efficiency is also limited, as it depends on the cross-section for photoabsorption.

There are many other types of ionization sources, some of them have advantages in the sample introduction techniques. For instance MALDI allows to desorb, vaporize and ionize solid samples directly, and ESI eases the analysis of liquid samples with the IMS.

### 1.3.1.2 IMS signal characteristics

#### **Reactant ion peak, proton affinity and charge competition**

When an indirect ionization source (such as radioactive or corona discharge) is used, the reactant ions generated from the supporting atmosphere will travel alongside the sample ions through the drift tube. If no sample has been introduced, reactant ions will travel alone. These ions (typically hydronium  $H_3O^+$  or ammonia  $NH_4^+$ , depending on the supporting atmosphere) appear as one or two peaks in the IMS spectrum. Figure 1.2 shows two IMS spectra measured with

a radioactive IMS. The blank sample shows the Reactant Ion Peak (RIP), while the other sample (consisting of a mixture of acetone and ethanol) shows several peaks corresponding to reactant ions and to the sample ions.

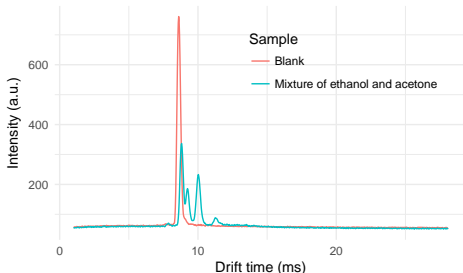
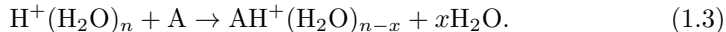


Figure 1.2: IMS spectra measured with a 63 Nickel radioactive source of a blank and a mixture of ethanol and acetone.

The ion formation mechanisms are not yet fully understood. The general idea in a common water-chemistry scenario is that beta radiation ionizes water molecules forming the reactant ions  $\text{H}^+(\text{H}_2\text{O})_n$  and  $\text{O}_2^-(\text{H}_2\text{O})_n$ . In positive ionization mode, the analytes with higher proton affinity than water will compete to get protons transferred to them according to reactions like the one shown at equation (1.3) (Eiceman et al., 2013). Similarly, in negative ionization mode the analytes will compete depending on their electronegativity.

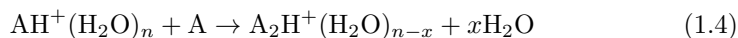


The amount of reactant ions available is limited, so in presence of multiple analytes those with lower proton affinity will not be ionized. This charge competition effect causes important non-linearities in complex mixture scenarios, as the sensitivity of the IMS to a specific analyte will decrease if that analyte is mixed with another analyte of higher proton affinity. Looking at the bright side, the use of reagent gases and dopants in the supporting atmosphere brings many possibilities to interfere with the ionization chemistry and to control the ion mobility of the formed ions. More information on this topic is available at (Puton et al., 2008).

### Adduct formation and IMS non-linearities

Another characteristic of the IMS that introduces non-linearities is the adduct formation in the form of protonated monomers and dimers. We have seen how equation (1.3) controls the formation of ion protonated monomers of a given

analyte. When the vapor concentration of the analyte increases, a protonated dimer appears following equation (1.4), at the expense of both the RIP and the monomer peaks.



Equations (1.3) and (1.4) show the relations between the RIP the monomer and the dimer peaks. The evolution of the intensities of an analyte that increases and decreases its concentration is illustrated schematically on figure 1.3. When the analyte concentration increases, the monomer intensity increases at the expense of the RIP, up to the point that dimer clusters start to form. At this time, the dimer intensity starts increasing at the expense of both the monomer and the RIP. When the concentration of the analyte starts decreasing the dimer will decrease as well, the monomer may briefly increase to decrease again and the RIP will recover its initial intensity.

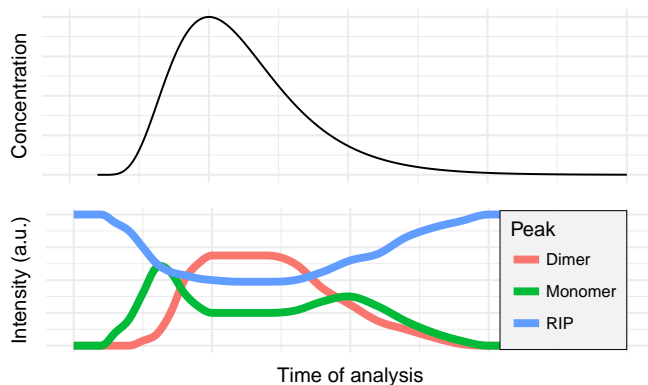


Figure 1.3: IMS RIP/monomer/dimer peak intensities example

If the concentration of the analyte is too high, the RIP may deplete. This is not desirable, as any quantitative calibration of the analyte in RIP depletion conditions will not hold. Moreover, such high concentrations of analyte in the ionization region can lead to diffusion of neutral molecules in the drift region. In these conditions, product ions form cluster ions with neutral adducts. These cluster ions have very short spans as they are easily broken in the drift tube. The formation and breakage of cluster ions in the drift tube is problematic as it increases the variability in the time it takes for the molecules to reach the detector. This produces broader peaks in the spectrum, centered on drift times that correspond to the weighted average of the times of all the cluster ions.

In summary, there are two causes of non-linearities in the IMS response, namely:

- Finite number of reactant ions that cause charge competition, when indirect ionization sources are used.
- Monomer and dimer formations that require multivariate non-linear methods for calibration of the concentration of analytes, even without sample mixtures.

Even if the high sensitivity and speed of analysis make IMS interesting for the analysis and detection of VOCs, the non-linearity issues present in the analysis of complex mixtures put IMS in a very challenging position for the analysis of biological samples. To overcome this, hyphenated analysis techniques are used to pre-separate analytes in mixtures, and avoid charge competition effects.

### 1.3.1.3 Other IMS variants

We have seen the operating principle of the most common type of ion mobility spectrometry and the main characteristics of their signals. It is worth mentioning that apart of the most conventional drift tube ion mobility analyzer there are other also common IMS technologies. For instance, there is the Field Asymmetric Ion Mobility Spectrometer (FAIMS) that applies higher electrical fields (10kV/cm) in a periodic asymmetric shape perpendicular to the ion movement. This wave makes ions drift and oscillate in an asymmetric way. FAIMS relies on the fact that ion mobilities are not constant in high electric fields to drag ions towards the lateral plates. As ions have different mobilities and different variations of mobility, by applying a compensation voltage on top of the drift wave we can achieve ion selectivity. A sweep on the compensation voltages will allow us to obtain a compensation voltage spectrum.

There is ongoing research interest in IMS miniaturization (Cumeras et al., 2012), (Kaye and Stimac, 2015) and even some miniaturized versions of FAIMS exist<sup>2</sup>. The interest in IMS miniaturization is related to the possibilities of coupling IMS with other analytical instrumentation. While in this thesis we will focus on Gas Chromatography – Ion Mobility Spectrometry, other hyphenated techniques such as Mass Spectrometry have been coupled to IMS. In a IMS-MS setup the IMS is used as a separation technique before further analyte identification, as recently reviewed in (Laphorn et al., 2013). Some more complex setups of GC-IMS-MS hyphenations have been used in explosive detection scenarios (Marr and Groves, 2003), although this triple hyphenation does not seem to be common in the literature.

---

<sup>2</sup><https://www.owlstonemedical.com/products/ultraFaims/>



### 1.3.2 Gas chromatography

Our interest in Gas Chromatography (GC) as an analytical technique emerges when we use it as a pre-separation technique for IMS, to palliate the selectivity issues and charge competition effects IMS has. We will describe the basics of the GC and briefly discuss GC-MS and GC-IMS hyphenations before moving to data analysis techniques.

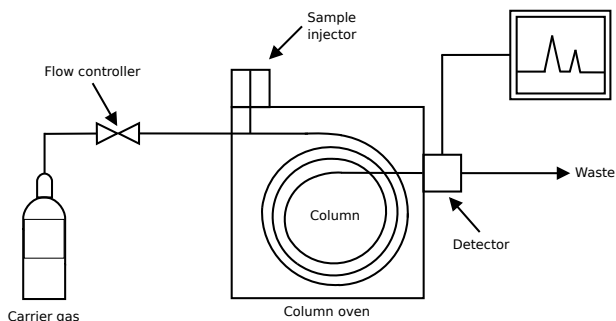


Figure 1.4: Gas Chromatograph schema. Wikimedia Commons / Public Domain

Figure 1.4 shows a simple diagram of a Gas Chromatography (GC) system. GC is the most common analytical technique for the separation of components in volatile gas-phase samples. As in all chromatographies, the sample is dissolved in a *mobile phase* that carries it through a different material called the *stationary phase*. The speed at which the analytes of the sample travel through the stationary phase depends on the physical and chemical properties of the analytes thus achieving the separation of components according to the time they need to pass through the stationary phase, the *retention time*. In GC, the sample is diluted in an inert or nonreactive gas (helium, hydrogen) that acts as the mobile phase. This is injected into a capillary column that is a typically long tube coated in the inside with a microscopic layer of a liquid or polymer. The analytes in the sample interact with the coating, eluting from the column at different retention times. Multiple parameters affect the performance of the separation, among them the selection of the carrier gas and flow, choosing the right coating of the column, the sample injection technique, and the column temperature. The chromatographic column is placed inside an oven of controlled temperature. At higher temperatures analytes elute faster and chromatographies take less time, however the separation of the analytes in the chromatographic spectrum is also lower. There is therefore a trade-off in the election of the temperature program, having the length of an experiment on one end versus the ability to discriminate

more analytes in the other end.

The GC as a separation method requires a detector at the output of the column. Flame Ionization Detectors (FID) are the most common, as well as Thermal Conductivity Detector (TCD) and Electron Capture Detectors (ECD). FID is suitable for the analysis of life sciences samples as it detects all organic compounds. Details on these and other detectors are available at (McNair and Miller, 2009, ch. 7).

Gas Chromatography offers good separation and high peak resolution so in some situations it can help in analyte identification. Towards that end and in order to standardize the differences between chromatographic column compositions and methods, Kovats Retention Indices or simply Retention Indices (RI) are often used to help identify organic compounds. The RI of an analyte that eluted at a given retention time is the result of applying a simple transformation with respect to the retention times of n-alkanes (that need to be analyzed). Further details are given in (Kováts, 1958) and (Nič et al., 2009).

When further identification is required, GC is coupled with Mass Spectrometry as a detector, that gives further analyte identification information. When a GC-MS hyphenated setup is used, the output of the chromatogram is represented by the Total Ion Count (TIC), that consists of the addition of all the mass fragments detected.

The TIC of the headspace of a human urine sample is shown in figure 1.5, revealing hundreds of compounds detected. This particular TIC shows a prominent baseline that increases at larger retention times. This baseline needs to be corrected, as it indicates some contamination from stationary phase in the chromatogram. The dynamic range of the peaks of the chromatogram ranges in 3 to 5 orders of magnitude, so precise detection of peaks even close to the noise level is required.

While GC offers good sample separation and large dynamic range, its combination with mass spectrometry allows us to obtain a fingerprint of the fragmentation of each of those peaks, making feasible the identification of the compounds in complex biological samples.

### 1.3.2.1 Gas chromatography – Mass Spectrometry

When further identification of the compounds in the sample is required, the use of mass spectrometry (MS) as a companion to gas chromatography is a golden standard for VOC analysis.

As analytes elute from the chromatographic column they reach the mass spectrometer. The mass spectrometer consists of an ion source, a mass analyzer, and

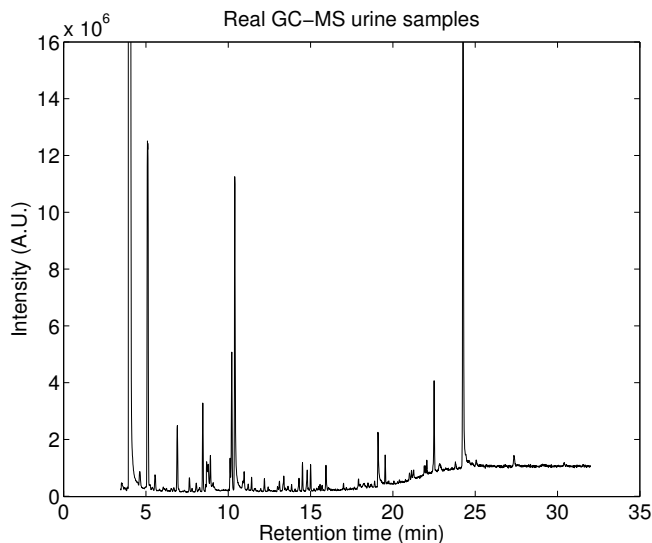


Figure 1.5: Total Ion Chromatogram of the headspace of a human urine sample. It presents a significant baseline and a large dynamic range.

a detector. While for liquid and solid samples ElectroSpray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI) are often used, for gas-phase samples Electron Ionization and Chemical Ionization are the most common. Ions are selected according to their mass/charge ratio ( $m/z$ ) in the mass analyzer. These mass analyzers can be based on several principles, such as a simple time of flight measurement under an electrical field, a quadrupole mass analyzer (where oscillating electrical fields are used to select the ion paths that will be detected), or Orbitrap (where ions are trapped in the orbit of a spindle shaped electrode that confines the ion and oscillate along the spindle axis. The frequency of the oscillation is recorded and it is related to the mass charge ratio of the ion so by Fourier transformation of the raw signal the mass spectrum is obtained).

Finally, the ions reach a detector that can be an electron multiplier based detector.

The range of prices and resolutions of Mass Spectrometry instrumentation are very broad. Several definitions of resolution are used in mass spectrometry. The IUPAC offers several definitions for MS resolution:

- 10% valley criteria: Given two adjacent peaks of equal height, compute the ratio  $\frac{m}{\Delta m}$ , where  $m$  is the  $m/z$  of the second peak and  $\Delta m$  is given as the

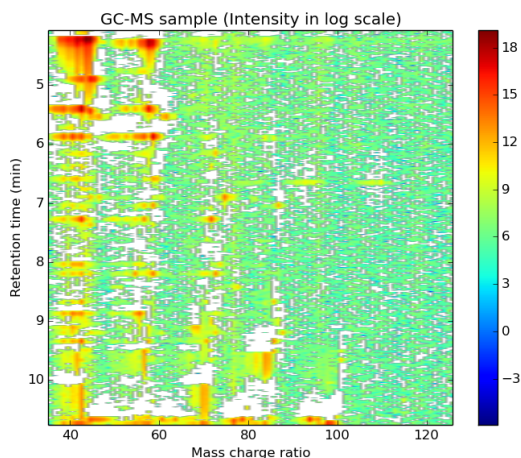


Figure 1.6: GC-MS of the headspace of a human urine sample.

distance between the two peaks, given that the valley between the two peaks reaches 10% of the peak intensity

- Peak width definition: Given a peak found at  $m/z$   $m$ , the resolution can be given as  $R = \frac{m}{\Delta m}$ , where  $\Delta m$  is defined as the width of the peak measured at a specific fraction of the peak height, that must be reported. Usual values for the fraction are 50% (that gives the ‘Full Width at Half Maximum’ or FWHM), 5% (that provides a definition technically equivalent as the 10% valley criteria) or 0.5%.

Another definition of the mass spectrometer resolution that is used but not considered by IUPAC is simply “unit resolution”. This is used in some quadrupole mass spectrometers to describe the ability to separate two consecutive integer masses.

There is a wide range of resolutions in mass spectrometers. Low resolution mass spectrometers (e.g. Thermo Fisher DSQ-II<sup>3</sup>) have a resolution of 1  $m/z$  unit in the 1-1000  $m/z$  range, while high resolution mass spectrometers have resolutions of  $R = 50000$  (FWHM, measured at 272  $m/z$ ) (Exactive GC Orbitrap GC-MS System<sup>4</sup>). Doing a rough comparison, DSQ-II would be able to distinguish mass 272 from mass 271 while the Orbitrap would be able to distinguish mass 272 from mass 271.99456, being more than a hundred times more precise. Therefore, high

<sup>3</sup>[http://www.thermo.com.cn/Resources/200802/productPDF\\_26943.pdf](http://www.thermo.com.cn/Resources/200802/productPDF_26943.pdf)

<sup>4</sup><https://www.thermofisher.com/order/catalog/product/0725510>

resolution mass spectrometers offer the possibility of determining the exact mass of the fragment ions of the molecules, providing better molecule discrimination and identification.

### 1.3.2.2 Gas chromatography – Ion Mobility Spectrometry

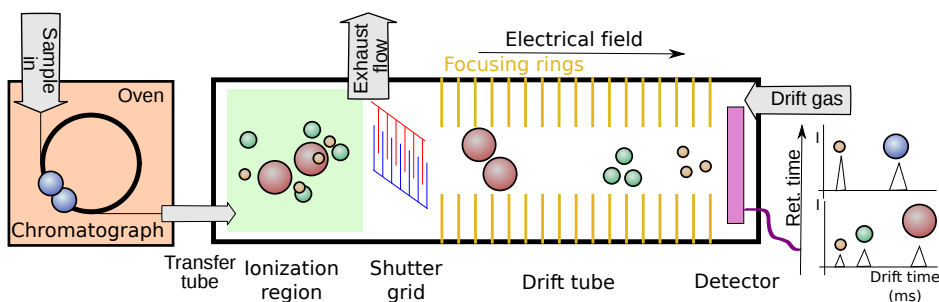


Figure 1.7: Gas Chromatography – Ion Mobility Spectrometer diagram

As we have already mentioned, the use of IMS for the analysis of complex samples requires some sort of sample pre-separation. The combination of GC with IMS as detector would ideally produce pure or nearly pure fractions of sample analytes and analyze each sample fraction separately. In reality, the high sample flow requirements of the IMS and its dynamics, as well as the desire for having fast chromatographies (as speed of analysis is one of the advantages of IMS) leads to the use of Multicapillary Columns (MCC). These columns are able to provide moderate separation in few minutes with higher sample flows, as they are made of multiple individual capillaries placed in parallel (Eiceman and Feng, 2009). The data obtained with MCC-IMS setups is characterized by broader peaks with higher degrees of coelution, that can separate simple mixtures but require advanced signal processing techniques for the analysis of more complex samples.

The pre-separation step provided by the MCC (a) helps discriminating analytes with similar IMS drift times if they present different elution times and (b) reduces the number of analytes simultaneously present in the ionization region, reducing the described charge competition effects.

Figure 1.8 shows an MCC-IMS sample of the headspace of olive oil. The peaks of this sample are much broader than the peaks of the TIC seen at figure 1.5 and there are several of them eluting at the same retention time. At a drift time close to 6ms one can see the RIP along the retention time and check that when

compound elutes from the column and gives a peak the RIP decreases accordingly.

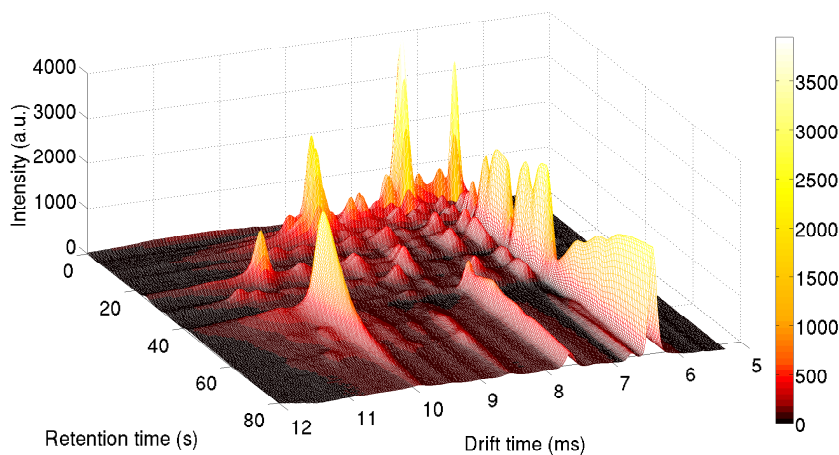


Figure 1.8: Region of a MCC-IMS olive oil sample. The Reactant Ion Peak (RIP) is observed at 6 ms. Multiple peaks on the same retention time indicate a strong co-elution.

In an equivalent way to how the Total Ion Count is computed on figure 1.5, MCC-IMS and GC-IMS two-dimensional samples can be projected to a single retention time axis. Instead of summing the intensities of all drift times for a given retention time, we can compute the RIP area and subtract it from the maximum, obtaining a non-selective figure of merit of the amount of charge that has been transferred to other ions throughout the retention time. This figure of merit is called the “Reverse RIP” and an example of it is shown on figure 1.9. The reverse RIP, compared to the TIC, presents much wider peak shapes and less selectivity, but it is useful for visualizing alignment results and in feature extraction techniques.

Similar MCC-IMS setups have been applied in the detection of gasoline components (Baumbach et al., 2003), determination of odd-flavors in foods (Márquez-Sillero et al., 2014) and bio-marker discovery applications in breath (Bödeker et al., 2008).

We are interested in the development of algorithms for MCC-IMS spectra because MCC-IMS can provide faster chromatographies than conventional GC-MS setups. Additionally, as MCC-IMS setups operate at ambient pressure, it is much easier to use them in out of the lab environments, such as point of care medical setups

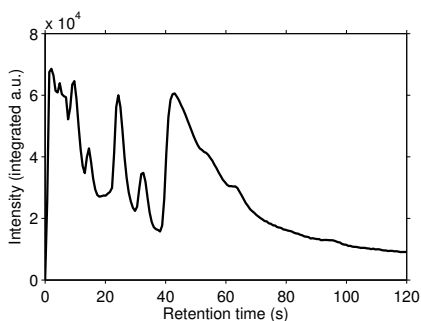


Figure 1.9: Reverse RIP. Analytes eluting from the column will show as a peak in the reverse RIP. The RIP is computed as the integral (from 6.26 to 6.6 ms) of each IMS spectrum and it represents the charge that has not been transferred to other analytes at a given retention time. The RIP’s maximum represents the total charge available. By subtracting the charge that has not transferred to the total charge, the reverse RIP is obtained.

and industries. This feature is concomitant to the lower cost of an MCC-IMS instrument compared to GC-MS instrumentation.

## 1.4 Data analysis for IMS and hyphenated instrumentation

With an overview on the kind of life sciences samples we want to focus on, and with the overview of the analytical instrumentation in the field, we have a solid base for the discussion on the existing algorithms for data analysis of hyphenated data, and its application to GC-IMS data. In this section we introduce the main data analysis techniques and challenges for IMS and hyphenated instrumentation, while in the following chapters these techniques are studied in more detail.

### 1.4.1 Preprocessing

Data preprocessing is the enhancement of raw data by the use of filters and in general signal processing techniques in order to remove noise and artifacts from the data and correct instrumental drift and baseline variations. The goal is to go from “raw data” (as it comes from the instrument) to “clean data”, ready for data analysis and modelling. The increase of quantity and complexity of raw

data makes it necessary to devote more time and efforts into data preprocessing to make sure that the information extracted from the samples later on in the data analysis pipeline is meaningful and of high quality. If preprocessing is not done carefully, artifacts and unwanted variations can be added to the raw data (instead of removed!) that can mislead further analyses (Engel et al., 2013).

Preprocessing is also necessary in order to be able to compare data measured from different measurements (for instance from two IMS), as well as to compare data from different laboratories. Choosing a preprocessing workflow is in general problem dependent, as it often depends on the experimental design to account for the confounding factors in the analysis. For instance, if an experimental design consists of several analysis batches, the preprocessing methodology used should account for that, reporting if batch effects are present or even trying to correct or minimize those batch effects in the cleaned data.

For IMS and GC-IMS data, the main problems that need treatment are random noise in the measurements, RIP detailing, baseline offsets, spectral misalignments and normalization.

#### 1.4.1.1 Denoising methods

IMS spectra as captured by the detector are typically noisy. The simplest denoising method usually applied is spectral averaging. This technique, consisting on simply averaging several spectra, is sometimes implemented at a firmware level in some devices, letting the user tune the number of spectra to average. This method is often applied in blocks, reducing the spectral sampling frequency by a factor of the number of spectra to average: For instance, if 32 spectra are being averaged and each spectrum is acquired in 21 ms, we will obtain an averaged spectrum every 672 ms. When this method is used in a GC-IMS instrument, this averaging is the limiting factor in the retention time sampling frequency.

This averaging technique is often combined with other preprocessing methods, such as digital filters or wavelets. The Savitzky-Golay filter (Savitzky and Golay, 1964) is often used for denoising. It is fast and simple to implement and can preserve the peak shape and area.

Another technique quite common is wavelet based denoising (Bader et al., 2008). This technique can be used to remove the noise and compress the signal by eliminating the first wavelet scales (removing high frequency components) and removing amplitude components below a threshold in the transformed space.



### 1.4.1.2 Baseline estimation methods

Baseline estimation techniques are often used to correct long term instrument contamination or degradation, as well as to correct for RIP detailing. They are essential for accurate peak area integration.

A simple and conventional way to estimate the baseline of a peak before the integration of its area consists of determining all the peak boundaries in a spectrum and fitting a soft curve to those points. That curve is the baseline. While the manual selection of peak boundaries and then fitting a curve to them to estimate the baseline of each peak is not an uncommon procedure, it is very expensive when the number of peaks in the signal increases (such as in complex biological samples) or when there is a large number of signals to analyse. Moreover, the analyst adds a subjective component to peak identification that depends on her/his expertise. For these cases, an automatic baseline estimation method is needed.

There are many automatic baseline estimation methods published, such as methods based on polynomial fitting (Salit and Turk, 1998), methods based on weighted least squares (Eilers, 2003), (Zhang et al., 2010), (Peng et al., 2010) or methods based on wavelets (Shao et al., 2003). Some of the methods require the user to define in advance regions without peaks to estimate the baseline. Other methods approach the baseline estimation iteratively, trying to detect and reject the regions with peaks that should not be part of the baseline.

The proposed method in this work consists of a modification of the Asymmetric Least Squares (ALS) baseline removal technique developed at (Eilers Paul H. C., 2005). We found that ALS technique suffers from bias in the presence of intense peaks (in relation to the noise level). These intense peaks are often found in GC-MS samples, as well as GC-IMS samples.

In chapter 2.2, a modification (named **psalsa**) to the asymmetry weights of the original ALS method is proposed to better reject large peaks above the baseline. Our method will be compared to several versions of the ALS algorithm using synthetic and real gas chromatography signals.

### 1.4.1.3 Alignment methods

After the sample noise has been reduced and the baseline removed from the data, alignment is the one major data preprocessing step left. Spectral alignment issues affect both IMS and GC instrumentation, as well as other analytical chemistry instruments (e.g. Nuclear Magnetic Resonance, Near InfraRed spectroscopy...)

Here we will focus on the description of the IMS and GC alignment issues, and several possible strategies to overcome them.

As described in section 1.3.1.1, on an IMS spectrum, minor changes to pressure and temperature conditions can cause variations of the ion mobilities, that shift peak positions in the spectra. These pressure and temperature changes can be partially corrected by converting the drift time to reduced ion mobilities. However, the correction is not perfect as flow variations and impurities can affect as well the mobility of the ions travelling through the drift tube.

As the reduced ion mobility is the main feature the IMS provides for analyte identification, these shifts in the peak positions need to be corrected in order to be able to compare spectra from several samples.

The GC misalignments have different causes: As mentioned in section 1.3.2, the chromatographic column degradation, as well as temperature and pressure variations affect the retention time at which molecules elute from the column. In order to compare several sample injections and be able to ensure that the retention times of two peaks correspond to the same analyte, these retention time variations need to be compensated.

In general, the alignment methods for spectral data can be divided in two groups, according to two possible strategies to follow: *Peak matching* and *Spectral warping*.

Peak matching is the approach used in LC-MS and GC-MS data analysis tools, such as MZMine (Pluskal et al., 2010) or PyMS (O’Callaghan et al., 2012). It consists on extracting features first from the samples (usually using a peak detection and a peak integration technique) and forming a list of peak tables, one table for each sample. This “peak picking” approach is followed by the “peak matching”, where all the peak tables are merged into a single table. In this peak matching, peaks from different samples corresponding to the same analyte are matched, and missing or noise values are used when a peak is not present in a specific sample.

Spectral warping keeps the whole spectra in further analyses. With this strategy, the retention time and/or drift time axes are warped in order to align the spectra from different samples so all the spectra are comparable. These techniques are often seen in spectroscopic data, such as Nuclear Magnetic Resonance, where the warp function is applied to the chemical shift axis to correct for pH fluctuations, among other issues. Many algorithms have been developed to warp in specific ways the axis according to instrumental knowledge, being *icoshift* (Savorani et al., 2010), (Tomasi et al., 2011), *Parametric Time Warping* (Bloemberg et al., 2010) and *Correlation Optimized Warping* (Nielsen et al., 1998) the most

popular techniques.

Both strategies have both advantages and disadvantages. Peak matching techniques require to keep a minimal amount of data from each sample (typically a list of peak positions and peak areas) making further data analysis faster and requiring less memory. However, they are very sensitive to peak detection and integration limitations: undetected peaks will end up being discarded, and partially overlapping peaks may be merged into a single peak with an erroneous peak area value. The peak matching algorithm must be properly validated, to prevent that peaks from different samples corresponding to different analytes are treated as a single analyte (peak merging error) or the opposite, that a single analyte appearing in multiple samples is not properly merged (peak splitting).

On the other hand, spectra profiling techniques keep the whole spectra for further data analysis. This strategy ensures that no feature or peak is discarded, no matter how small it is, and avoids the peak merging issue of the peak picking approach as there is no area integration. However, this strategy has a higher computational cost and it is very sensitive to the spectral alignment process: The warping functions used for the alignment of the spectra may introduce artifacts in the signal or may align mismatched peaks. These alignment issues are similar in nature to the ones found with the peak picking approach.

Conventional data acquisition software for GC-IMS data (e.g. LAV<sup>5</sup>) provides manual alignment methods, limited to linear distortions of the retention time and drift time axes. For each pair of samples (one of them being the reference), the user needs to select for the retention time axis and the drift time axis two coefficients corresponding to a linear relation as given by (1.5). This linear correction applied to the drift time axis is able to compensate for pressure and temperature variations, but it is unable to correct for non-linearities. Moreover, this approach is also time consuming and dependent on the skills and criteria of the user.

$$\begin{aligned}t'_{ret} &= a_1 t_{ret} + a_0 \\t_{drift} &= b_1 t_{drift} + b_0\end{aligned}\tag{1.5}$$

More complex techniques exist, which are able to warp the spectra either by compressing or expanding them, inserting or removing spectra segments, either processing each spectra as a whole or by pieces. Usually these techniques aim to maximize the correlation between the spectra, subject to constraints related to how (and how much) can the spectra be distorted.

---

<sup>5</sup><http://www.gas-dortmund.de/index-gas.php?lan=1&spath=463>

In 2.3 we will discuss alignment strategies and their corrections both in the retention time axis and in drift time axis, proposing a protocol for the correction of the retention time in GC-IMS samples across a several months long study.

#### 1.4.1.4 Normalization

Sample normalization techniques are used to compensate two effects that present similar consequences (a) variations in the instrument sensitivity and (b) variations in sample weight, volume or concentration. These effects need to be considered on each study, and are fundamental in order to compare data from several studies.

If an instrument presents fluctuations in its sensitivity, and the samples are not properly randomized in the experimental design, there is a danger that samples representing a specific condition present larger/smaller values due to the instrument fluctuations, confounding the results. One or more internal standard analytes, not expected to be found in the sample, and spiked before the analysis can be used to compensate for those variations in instrument sensitivity (Sysi-Aho et al., 2007). External standards, measured in samples interleaved in the experimental design can be used to monitor the instrument sensitivity across the study. On the other hand, unwanted variations in sample concentration are also susceptible of being corrected through normalization. For instance in a urine analysis, the urine concentration can vary from subject to subject according to the amount of water they have drunk. A traditional and simple approach to compensate for this effect is to use data scaling methods, such as total area normalization where each sample is normalized to present the same total area under the assumption that the overall concentration of all the analytes should be similar. When this approach is used, any artifact peak coming from the instrument can distort the total area, and therefore distort all the peak areas after the normalization, unless they are specifically excluded (Shellie et al., 2005). An alternative to this coarse normalization method is to use a fine grained approach and normalize to a specific analyte or a small number of analytes known to present a low variability.

While this thesis does not focus on normalization methods, a research article with related discussion has been published at (Cominetti et al., 2016), where the proteins of human plasma samples were analyzed. In that work, a bovine protein was used as internal standard for quality control of the instrumental variability and we discussed what human plasma proteins are good candidates to be considered cross-study protein standards.

The use of specific analytes for normalization purposes is platform dependent and domain dependent, being in most cases an open problem without a widely adopted consensus, despite the efforts.

### 1.4.2 Feature extraction: Peak deconvolution

One of the issues of the peak picking approach described previously is that the determination of peak boundaries may be hindered by the partial overlapping of peaks corresponding to the same analyte. This happens often in IMS spectra, where the broad shape of the peaks makes it likely for several analytes to be at least partially overlapped in the spectra. When an MCC-IMS instrument is used, the peak overlapping is present as well in the retention time axis, as the co-elution is a prevalent phenomenon in MCC, as described in (Eiceman et al., 1995), (Baumbach, 2009) and (Eiceman and Feng, 2009). To overcome this instrumental limitation, there are Blind Source Separation techniques (BSS) also named in chemometrics “resolution techniques”. These techniques commonly applied to data from hyphenated methods, for their capability to deconvolve a matrix of spectra into a matrix of pure components (or pure spectra) and their concentration profiles.

In IMS samples, the compounds’ original concentration profiles and pure spectra can be deconvolved from the sample using BSS techniques, being the family of Multivariate Curve Resolution methods the most popular (Pomareda et al., 2010). In this thesis, we propose a blind source separation technique for MCC-IMS data. Direct application of MCR techniques to full MCC-IMS data typically fails to resolve co-elution due to the complexity of the data and to the global noise, which hinders the detection of weak but significant peaks. The typical approach in this case is the manual selection of the retention time window where the co-elution appears and the application of MCR in this data subset. However, few individual peaks can be isolated in the total chromatogram and mostly very broad peaks are observed. To deal with this complexity, we propose an automatic manner to investigate co-elution across the whole chromatographic axis.

The proposed method is able to detect and recover compounds in adverse co-elution conditions and reject spurious spectra with no physical meaning in an unsupervised manner. This is described in detail in chapter 3 and it is implemented in the developed toolbox.

### 1.4.3 Outlier detection and lab quality control

During a whole study, ranging from the sampling collection to the data analysis, many things can go wrong. Data analysis techniques must do their best to detect and prevent errors from spreading and damaging the whole study. This is specially relevant in biomarker discovery applications, where one of the major current issues is the poor translation of prediction models to clinical practice,

due to lack of reproducibility in the studies among other factors as mentioned in (Ioannidis and Khoury, 2011) and (Xia et al., 2013). The detection of outlier samples and confounding batch effects as well as their rejection is important in order to avoid undesired confounding factors that may lead to false discoveries. As described in (Ioannidis and Khoury, 2011), this issue applies not only to metabolomic studies but in general to biomarker discovery studies in -omics.

A sample can be an outlier for several reasons, usually either due to a biological condition or a problem in the sample acquisition (for instance having a urine sample excessively concentrated or diluted, or with a patient not complying with the experimental protocol), or instrumental issues (for instance injection malfunctions, chromatographic column contamination, strong spectral misalignments). A desirable data analysis protocol will include methods to detect outliers and reject them as well as to study their distribution across typical confounding variables, such as the sample collection day, the laboratory injection order, or the plate where the samples were stored. A large number of outliers on a given day or plate may suggest that the whole day or plate should be discarded or if possible corrected, and further investigations may tell the reasons behind the issue.

Apart from rejecting the outlier samples from a study, an unsupervised exploration of the samples can reveal an intrinsic structure, such as samples grouped in clusters. These clusters can be desired (e.g. if they are related to a condition or feature we want to discriminate) or undesired (e.g. related to an instrumental drift). In some cases clusters are unavoidable (e.g. gender or age differences in a biomedical study).

As an example, work in the characterization and quality control measures of a biomedical study can be seen in (Cominetti et al., 2016), where we presented a workflow for proteomic biomarker discovery in human plasma samples. In that work, the impact of instrumental confounders, such as the plate or collection center is assessed, as well as common standard clinical variables that typically affect results.

#### 1.4.4 Classification and Regression

Once either a) a table of peak areas, b) deconvolved concentration profiles or c) clean spectra profiles are obtained, common statistical tools or machine learning algorithms can be used to characterize or build a model able to answer the hypothesis under test.

If we aim for the prediction of a class (for instance a medical condition or whether or not the quality of the sample is acceptable), a classifier will be used. On the

other hand, if we aim for the prediction of a continuous feature, for instance a concentration of some analyte, we will train a regression model. In any case, all these techniques are based on assumptions of the underlying data, such as data normality, linearity, or the parameters that need to be set in order to train the model, and have their own limitations and requirements.

For instance in classification problems, both linear and quadratic discriminant analysis classifiers (LDA, QDA) assume that the underlying sample distribution is a multivariate Gaussian and easily overfit when the number of features is larger or even in the order of magnitude of the number of samples. For this reason they are often combined with a Principal Component Analysis to reduce dimensionality (Garrido-Delgado et al., 2012). Other methods such as Partial Least Squares - Discriminant Analysis are able to cope with a larger number of features, as they work by projecting the samples into a linear subspace of latent variables that maximize the covariance of the data with the class. This dimensionality reduction makes it very suitable for full spectra profile processing, typical of spectroscopic data (Worley and Powers, 2012), (Griffin, 2003).

In the applications chapter (section 4.1), we will show how a PLS-DA can be used to discriminate qualities of olive oil and how to validate the model to prevent overfitting.

### 1.4.5 Validation

As it has been mentioned before, one of the main issues with -omics studies is often a poor reproducibility of the results, that hinders the translation to the clinics. A way to overcome this is to use strict validation standards able to minimize the chances of false positives in the results.

A typical cause for the poor reproducibility is the overfitting of the model to the dataset used for training it, or in other words, that the trained model is not able to generalize to new samples. The classification and regression models have parameters that control their complexity. A too simple model won't be able to fit the data properly while a too complex model will fit both the data and the noise, and won't be able to generalize to new samples properly.

Internal validation techniques – such as K-Fold, Random Subsampling, or Leave One out – can be used in combination with a metric of the performance of the model – such as the classification rate, the area under the receiver operating characteristic curve, or the root mean square error – to estimate the optimal model parameters. The operating principle behind those techniques is the same: the dataset is partitioned into a training subset and an internal validation subset,

then the model is trained with the former and tested on the latter, sweeping through the model parameters. This procedure is repeated with several partitions, and the model with the best average score is chosen as the optimal one.

Finally, external “blind” samples that have not been used to train the model should then be used to assess the actual performance of the optimal model.

This procedure can be repeated multiple times, obtaining multiple “final” models. This approach is called “double cross-validation” and it is described at (Smit et al., 2007). Each of the “final” models will provide an estimation of its performance that can give us an idea of the distribution of the performance of the models to solve our problem. Ideally, the models should report similar performances and, if a model similarity metric can be defined, be similar among them.

Additionally, to discard that the results could have been obtained by chance, it is possible to repeat the analysis multiple times with permuted labels, obtaining a distribution of null performances. We should ensure that the performance of our trained model is significantly higher than the randomly obtained performances.

This thesis follows these validation procedures to ensure that the results are reliable and offers this validation techniques in the developed toolbox.

## 1.5 Objectives

This thesis focuses on data analysis methods for processing samples, measured using Ion Mobility Spectrometry (IMS) as detector with Gas Chromatograph as a pre-separation technique.

More specifically, this thesis aims to:

- Study algorithms and techniques for data processing of analytical instrumentation, with a special focus on hyphenated instrumentation and including IMS, GC-MS and GC-IMS.
- Adapt and develop algorithms for preprocessing GC-IMS samples. In particular reliable baseline estimation methods. This is addressed in section 2.2, where a baseline estimation method is proposed and benchmarked to state of the art alternatives using both simulated and real data.
- Characterize with a study the misalignments of GC-IMS samples, both in retention time and drift time, proposing a method based on monotonic cubic splines to correct it. This is addressed in section 2.3.
- Provide a feature extraction algorithm for GC-IMS data. This algorithm is based on the application of Multivariate Curve Resolution – Alternating



Least Squares in a sliding window across the chromatographic retention time and is described in Chapter 3.

- Apply the explored techniques in biological complex data, such as olive oil headspace samples (for quality control and fraud prevention applications) measured with GC-IMS, and urine headspace samples (for biomarker discovery applications), measured with GC-MS. This is described in Chapter 4.
- Offer the algorithms developed in an open source toolbox, that leverages on existing machine learning algorithms for further flexibility, gathers common techniques, and can provide a complete solution for the data analysis and improving the ecosystem of open source data analysis resources.

## Chapter 2

# Preprocessing

As explained in the introduction, data preprocessing techniques are used for enhancing raw data, removing noise and artifacts so further extracted information is as clean as possible and has nice properties for data modelling methods, that take care of building classification or regression models.

Both IMS and GC-IMS, as most analytical techniques, need preprocessing before further peak extraction, especially when dealing with analytes in low concentration, close to the detection limits. While there are many data processing resources for preprocessing analytical chemistry instrumentation — XCMS (Smith et al., 2006), MZMine (Pluskal et al., 2010)... — the number of open source tools for IMS and GC-IMS data preprocessing remains limited (IPHex (Bunkowski, 2012)) and not widely used, as it is common to rely on privative solutions provided by instrument manufacturers, such as VisualNow<sup>1</sup>) or LAV<sup>2</sup>. These tools work well with data from their instruments, but have limited possibilities for being extended with new algorithms.

In this chapter, we will deal with denoising, baseline correction and spectral alignment. Among this issues, more emphasis is given to the last two, as it is where some of this thesis contributions are made.

---

<sup>1</sup><http://www.bs-analytik.de/en/products/software-vocan-visualnow.htm>

<sup>2</sup><http://www.gas-dortmund.de/index-gas.php?lan=1&spath=463>

## 2.1 Denoising

Noise removal algorithms are often needed in analytical instrumentation analysis, with both very simple and effective solutions as well as more complex ones.

Often, instrument firmware has the ability to perform some denoising, for instance by averaging a number of consecutive samples. This approach can reduce the noise in the samples at the expense of reducing the sampling frequency of the instrument. For instance, the Gas Detector Array (GDA) IMS instrument from Airsense reports the median of several spectra, and the GC-IMS instrument from FlavourSpec offers the user to tune how many spectra should be used for averaging. If a single IMS spectra is acquired in 21 ms and the instrument is averaging 32 spectra, the sampling period in retention time will be of 672 ms. Increasing the number of spectra averaged will further reduce the resolution in retention time, while reducing that number will provide more spectra but noisier.

Besides the denoising used in instrument firmware, signal processing filters are often used to enhance the signal to noise ratio, as well as filters applied in domain transformations such as Fourier transforms or more recently Wavelet transforms (Wentzell and Brown, 2000).

For IMS and GC-IMS data, (Bader et al., 2008) recently used a Daubechies 8 wavelet transformation for compressing and denoising GC-IMS chromatograms, removing components with either high-frequency or small-amplitudes. This approach was also used in (Szymańska et al., 2015), where they combine the wavelet denoising with a mask selection, that discards from further analysis regions without information. Previously, this wavelet approach was formally presented by (Donoho and Johnstone, 1994) and optimized for chemometrics applications by (Pasti et al., 1999).

Digital filters are the other typical approach for denoising IMS and GC-IMS samples as used in (Karpas et al., 2012b) or (Bunkowski, 2012). This approach is common as well in other analytical techniques such as GC-MS (Hoffmann and Stoye, 2012). The most common digital filters used for denoising are either median or average moving filters sometimes followed by a Savitzky-Golay filter (Savitzky and Golay, 1964). In (Guamán Novillo, 2015), more complex solutions are developed for specific IMS instruments with very low signal to noise ratio with some overimposed periodic noise components. These more complex solutions used either Principal Component Analysis (PCA) or Independent Component Analysis (ICA) to remove the components associated to undesired noise. While those complex strategies were proven to significantly increase the signal to noise ratio, they are not vital, assuming proper electronic insulation of the instrument.

This thesis uses the existing denoising methods, as they cover the denoising needs for cleaning noise from GC-IMS and IMS samples.

## 2.2 Baseline estimation

As mentioned in the introduction, baseline estimation techniques are required to correct long term instrument contamination or degradation, and for accurate peak area integration.

There are many ways to correct baseline issues. The most rudimentary one is to manually select the peak boundaries and fit a curve to them to estimate the baseline of each peak. However manual baseline estimation is very expensive when the number of peaks in the signal increases (such as in complex biological samples) or when there is a large number of signals to analyse. Moreover, the analyst adds a subjective component to peak boundary identification that depends on her/his expertise. For this cases, automated baseline estimation methods are needed.

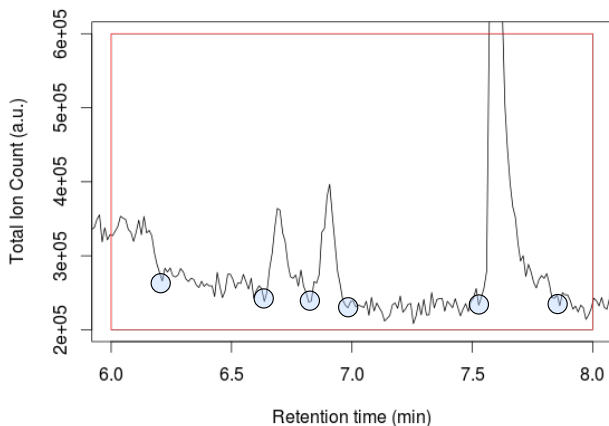


Figure 2.1: Region of GC spectra with manual peak boundaries marked

There are a wide range of methods for baseline estimation. Some methods are based on simple polynomial fitting (Salit and Turk, 1998). Other methods estimate the baseline iteratively, trying to ignore peaks which do not belong to the baseline (Gan et al., 2006). There are also tailored methods, for instance those

designed to perform well with samples that present a sparse number of peaks, where most of the intensities have baseline values, and use a low quantile of the intensity distribution to estimate the baseline (Bunkowski, 2012). There are also baseline estimation methods based on simple non-parametric techniques such as Asymmetric Least Squares (ALS) that have been recently gaining popularity (Eilers, 2003), (Zhang et al., 2010), (Peng et al., 2010).

In this section, we explore several baseline estimation methods. We focus on Asymmetric Least Squares, and some of its derivations, and we propose a modification of ALS tailored for spectral-like data, like GC, IMS and GC-IMS samples. We show how the ALS technique suffers from bias in presence of intense peaks (high intensity compared to the baseline). Our method, named **psalsa** and presented at (Oller-Moreno et al., 2014), improves the rejection of those large peaks, by modifying the asymmetry weights of the original ALS method. We benchmark its performance with respect to other methods, using both synthetic and real chromatographic data. Our proposal improves existing solutions both by providing more accurate baseline estimations and by being more robust to parameter variations (so less parameter tuning is required). The **psalsa** method is applied in chapter 4 (Applications), both in GC-IMS and GC-MS samples.

## 2.2.1 Data description

Two Gas Chromatography datasets are used to compare the different methods: On the one hand, a synthetic dataset offers the possibility to *objectively* assess the performance of the different methods, as we know the real baseline added to the synthetic signal and therefore we can compute the error of the different baseline estimations. On the other hand, a real dataset lets us check how the different methods perform on *real world* samples, which inevitably are more complex than synthetic chromatograms.

### 2.2.1.1 Synthetic dataset

A dataset with  $N_{\text{synth}} = 100$  samples was generated. Each synthetic chromatogram lasted 30 minutes long with a sampling frequency of 2 Hz. Each sample was the combination of three components: a **baseline**, some **random noise** and a **signal** made from the addition of several peaks. Figure 2.2 shows an example of one of those synthetic chromatograms.

Each of the components of the synthetic chromatograms was designed to obtain chromatograms similar to the ones found on the real dataset (described below).

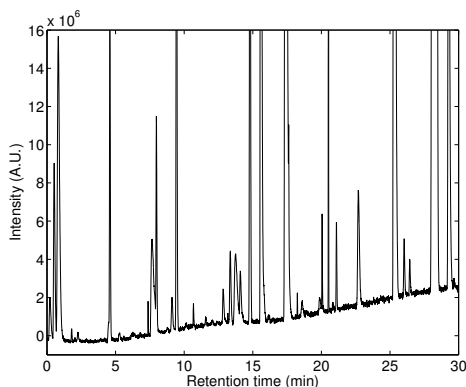


Figure 2.2: Synthetic chromatogram

The chosen criteria described below is also similar to the criteria followed by the NIST, in their chromatogram simulator<sup>3</sup>.

### Peak model

In order to generate the signal, several peaks are generated and placed randomly on the signal. A peak density of 0.25 peaks/s is chosen giving a total of 450 peaks/sample.

Peaks are modeled following a Generalized Exponential (GEX) function. The generalized exponential function (Felinger, 1998) is an empirical peak model that has been used successfully to describe chromatographic peaks (Vaidya and Hester, 1984), taking into account factors such as peak shape and peak asymmetry.

The GEX model is represented in figure 2.3 is given by:

$$f(t) = h \left( \frac{t - t_0}{t_m - t_0} \right)^{b-1} \exp \left\{ \frac{b-1}{a} \left[ 1 - \left( \frac{t - t_0}{t_m - t_0} \right)^a \right] \right\} \quad (2.1)$$

with  $a > 0$  and  $b > 1$  are constants,  $h$  is the peak height,  $t_m$  is the location of peak maximum and  $t_0$  is the time where the peak starts emerging from the baseline.

Peak model parameters are sampled from different probability distributions, with parameters in empirically reasonable values to obtain a synthetic dataset similar

<sup>3</sup><https://www.nist.gov/services-resources/software/simulated-chromatographic-data>

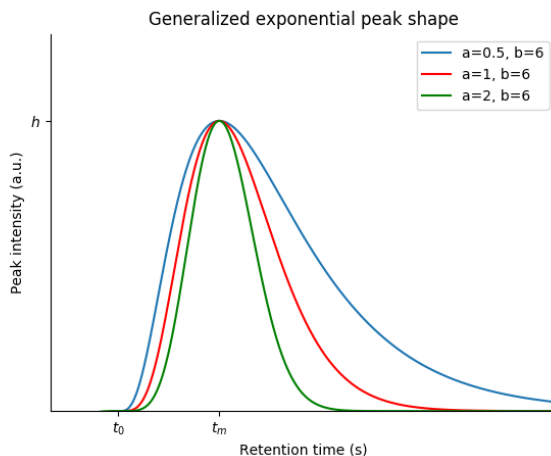


Figure 2.3: Generalized exponential peaks

to the real dataset:

- $a$ : Uniform distribution with  $\min = 0.5$  and  $\max = 2$ .
- $b$ : Uniform distribution with  $\min = 5$  and  $\max = 8$ .
- $h$ : LogNormal distribution with  $\mu = \log(400)$  and  $\sigma = \log(200)$ .
- $t_0$ : Uniform distribution in the retention time range.
- $t_m$ :  $t_0 + 2 + \text{Poisson distribution of } \lambda = 4$

### Baseline model

The baseline is generated following a combination of several contributions. The *ArcTan* factor contributes to the baseline by increasing it at larger retention times, in a similar way to slight column bleeds. The *Linear* and sinusoidal contributions simulate slower fluctuations.

$$\begin{aligned}
 b(t) &= \text{ArcTan}(t) + \text{Linear}(t) + \text{Sinusoidal}(t) \\
 \text{ArcTan}(t) &= A_{\text{low}} + \frac{2(A_{\text{high}} - A_{\text{low}})}{\pi} \cdot \arctan\left(\frac{\pi \cdot (t - t_0)}{t_r}\right) \\
 \text{Linear}(t) &= mt + n \\
 \text{Sinusoidal}(t) &= A \sin(2\pi f \cdot t + \varphi)
 \end{aligned} \tag{2.2}$$

The parameters for each baseline contribution are chosen from random uniform distributions in the following ranges:

- ArcTan:  $A_{\text{low}} \in [2, 3] \cdot 10^5$ ,  $A_{\text{high}} \in [1, 1.5] \cdot 10^6$ ,  $t_0 \in [1100, 1300]$ ,  $t_r \in [300, 700]$
- Linear:  $m \in [3.5, 6] \cdot 10^5$ ,  $n \in [4, 7] \cdot 10^5$
- Sinusoidal:  $A \in [5, 30] \cdot 10^4$ ,  $f \in [0.9, 1.4] \cdot 10^{-3}$ ,  $\varphi \in [-\pi, \pi]$

### Noise model

Gaussian noise with  $A = 100 + 200t$ ,  $\mu = 0$  and  $\sigma = 400$  has been added to the signal. The amplitude increases with the retention time to simulate the fact that the end of the chromatogram is more noisy than the beginning.

#### 2.2.1.2 Real samples

Chromatograms from a GC–MS dataset of human urine headspace samples were used to test the proposed algorithm. Figure 2.4 shows samples from this dataset, notice the large dynamic range on the  $y$  axis showing peaks orders of magnitude larger than the rest of the signal.

Samples were analysed at the PCB (Barcelona Scientific Park) premises, using a gas chromatograph – mass spectrometer (Focus GC–DSQ II) from Thermo Scientific equipped with a quadrupole analyser and an electron multiplier detector. The capillary column used was DB-624 (60 m  $\times$  0.32 mm i.d.) coated with 6 % cyanopropylphenyl 94 % dimethylpolysiloxane (film thickness 1.8  $\mu\text{m}$ ). The temperature program of the chromatographic oven began at 60 °C (2 min) ramped to 220 °C at 8 °C  $\text{min}^{-1}$  and held for 5 min. The injection port was maintained at 220 °C throughout the experiments.

### 2.2.2 Baseline estimation method description

In (Newey and Powell, 1987) Asymmetric Least Squares (ALS) was introduced in order to construct statistical tests for homoskedasticity, applying them to Econometrics. Much later, Eilers et al. applied ALS for baseline estimation in connection to Parametric Time Warping alignment (Eilers, 2004), and presented it in detail (Eilers Paul H. C., 2005). Recently, a modification of the ALS algorithm named *airPLS* was presented (Zhang et al., 2010), improving the weights of the original ALS method. Additionally, J Peng et al. (Peng et al., 2010) presented a different improvement to the original ALS method focusing on baseline estimation with multiple samples.



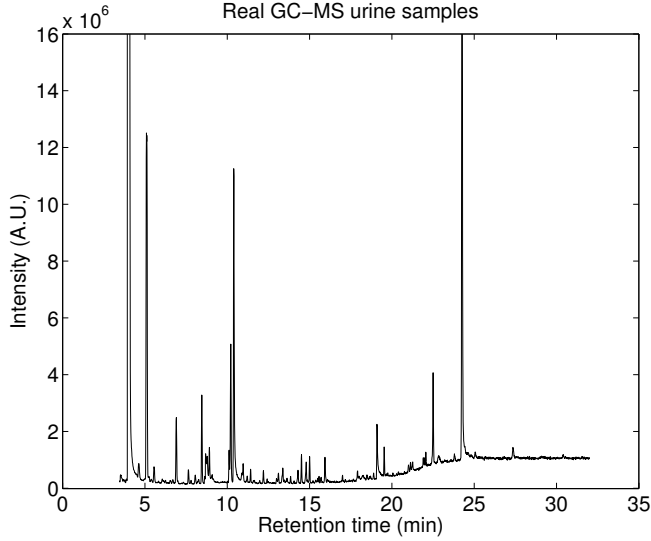


Figure 2.4: Real urine samples

### 2.2.2.1 Original Asymmetric Least Squares

Given a signal  $y$  of length  $m$ , ALS aims to estimate a signal  $z$  smoother than  $y$  but still similar to it. ALS proposes a model-free cost function given by:

$$S = \sum_i d_i^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (2.3)$$

where  $d_i = y_i - z_i$  are the residuals of the estimation and  $\Delta^2 z_i = z_i - 2z_{i-1} + z_{i-2}$ .

The first term in  $S$  accounts for the *fidelity* from  $z$  to  $y$ , while the second term imposes *smoothness* to  $z$ . Smoothness is controlled by parameter  $\lambda$ , usually chosen between  $10^2 \leq \lambda \leq 10^9$ . The cost function can be generalized by introducing weights  $w$ :

$$S = \sum_i w_i d_i^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (2.4)$$

These weights  $w$  are introduced so as, if properly defined, will be able to reject penalizations to the cost function produced by regions where the signal is above the estimated baseline (i.e. peaked regions).

The proposed definition of  $w$  is based on a parameter  $p$  which is usually chosen as  $0.001 \leq p \leq 0.1$ :

$$w_i = \begin{cases} p & \text{if } d_i > 0 \\ 1 - p & \text{otherwise} \end{cases} \quad (2.5)$$

As one can see from the definition of  $w_i$  and the values of  $p$ , regions where the signal is placed above the baseline will have a much smaller contribution to the penalty.

Minimization of equation (A.2) leads to:

$$(W + \lambda D' D) z = W y \quad (2.6)$$

where  $W = \text{diag}(w)$  and  $D$  being the difference matrix:  $Dz = \Delta^2 z$ . As there is no model imposed on  $z$ , there will be  $m$  equations forming a sparse system, where only the diagonal and two sub-diagonals above and below it are non-zero.

A solution to eq. (A.2) can be found by iterating. Given an initial set of weights  $w_i = 1$ , an initial estimation for  $z_i$  can be computed. From  $z_i$ , weights are computed and used to get a new estimation for  $z$ . Less than 20 iterations are needed for a proper estimation of  $z$ .

According to (Eilers Paul H. C., 2005), a proper value for  $p$  may be validated by considering the histogram of the residuals  $d$ , so as the noise components are normally distributed near zero and peaks are represented in the histogram as a positive asymmetric component. The right value for  $p$  will produce a baseline that cuts the noise instead of fitting below or above it.

One limitation of the ALS method appears when the chromatogram presents a large dynamic range, with peaks of very large intensity. In this case, the estimation of the baseline given by ALS will not converge to the actual baseline. If the ALS algorithm converged to the actual baseline, the  $d_i^2$  term below peaks with large intensity would be very large as well. In order to have a minimum in the cost function (convergence), either (a)  $p$  is chosen small enough to compensate for the large  $d_i^2$  or (b) the *smoothness* term is so large that the lack of fidelity driven by the  $d_i^2$  penalty is not relevant in that area (i.e. a large value of  $\lambda$  is chosen). In the case of (a), outside of the large peak areas, the small  $p$  values would force the convergence of the baseline to be *below* the noise, instead of crossing through the noise. Therefore, the rest of the peak areas will not be properly estimated. In the second case, the large value of  $\lambda$  would not allow the baseline to fit to baseline variations properly. Often, the best election of  $p$  and  $\lambda$  in these cases requires a

trade-off between a proper fitting of the larger peaks or fitting the smaller peaks. The result is a baseline estimation that penetrates inside the larger peaks. The baseline obtained in several iterations in one of such cases is seen in figure 2.5.

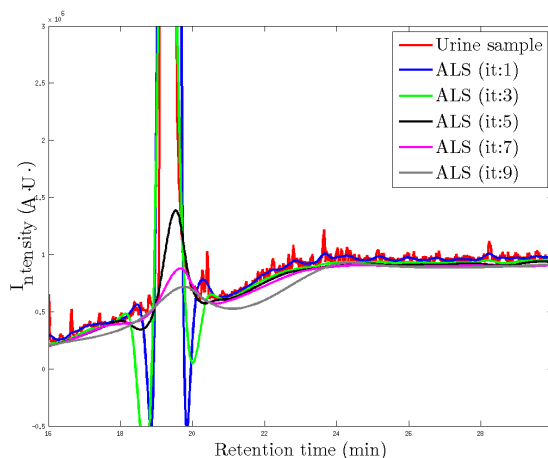


Figure 2.5: ALS fit in successive iterations. The final iteration still penetrates inside the large peak located at 19 min while it is fitted below the noise level (and not crossing it) in the rest of the chromatogram.

### 2.2.2.2 airPLS correction

In (Zhang et al., 2010), the authors proposed an improvement to the definition of  $w$  with two objectives: To remove the parameter  $p$ , simplifying the usage of the algorithm; and to improve the quality of the estimation by adapting the weights depending on the distance from the signal to the baseline.

The definition of the weight vector  $w$  for *airPLS* is as follows:

$$w_i = \begin{cases} 0 & \text{if } d_i > 0 \\ \exp\left(\frac{-t \cdot |d_i|}{\sum_{d_i < 0} |d_i|}\right) & \text{otherwise} \end{cases} \quad (2.7)$$

where  $t$  is the current iteration. With this definition of weights, regions of the signal where the signal is above the estimated baseline are ignored at the next

iteration. For the rest of the weights, the further the signal is from the baseline the least it contributes to the penalty.

Having the current iteration  $t$  in the exponent forces the weights to be smaller on each iteration, making more significant the *smoothness* term as iterations go on.

The criteria set by *airPLS* to stop iterating is given by either a maximum number of 20 iterations or by:

$$\sum_{d_i < 0} |d_i| < 0.001 \sum_{\forall i} |y_i| \quad (2.8)$$

The featured *airPLS* version 2.0 for MATLAB was used as the reference implementation. In this version, a  $p$  value is used to set the weights of points found at the beginning and at the end of the spectra as the adaptation of the weights does not give good estimates close to the signal limits.

### 2.2.2.3 Proposed method: psalsa

We propose a different definition for the weights much more similar to the original ALS algorithm. However, we define an adaptive value for the weights depending on the residuals as follows:

$$w_i = \begin{cases} p \cdot e^{-\frac{d_i}{k}} & \text{if } d_i > 0 \\ 1 - p & \text{otherwise} \end{cases} \quad (2.9)$$

The difference with respect to the original ALS method is on the positive residuals, where  $p$  is pondered by  $\exp\left(-\frac{d_i}{k}\right)$ . Peak regions will show large residuals getting smaller weights, whereas noise regions will present small residuals and weights close to  $p$ . This approach gives an additional parameter named  $k$  that controls the exponential decay of the weights. This parameter can be set to 5% of the maximum intensity value. Note that by taking the limit  $k \rightarrow \infty$  we recover the traditional ALS method.

As the original ALS method does, the criteria used by *psalsa* to stop iterating is given by either a maximum number of iterations (usually 10) or when the residuals do not change of sign with respect to the previous iteration.

## 2.2.3 Results

### 2.2.3.1 Synthetic chromatograms

The three described methods were applied to the synthetic chromatograms, to obtain an objective evaluation of the baseline estimation. A sample of the optimal fitting results is shown at figure 2.6.

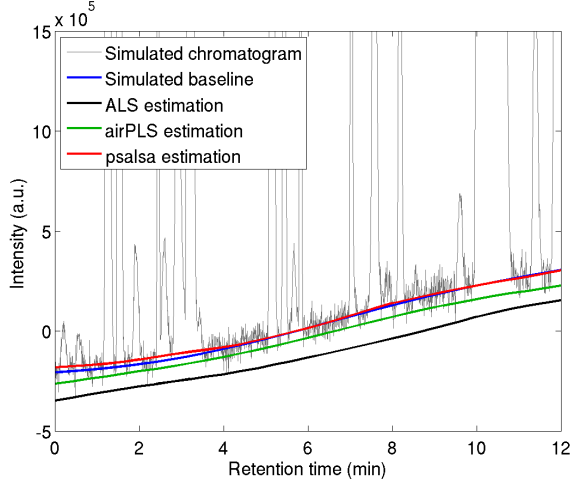


Figure 2.6: Region of a synthetic sample showing different baseline estimations.

In order to estimate the best parameters for each method, the parameter space was swept. For each sweep, the root mean square error (RMSE) was used as a figure of merit (defined in (2.10)). The RMSE values were averaged across samples, obtaining a global RMSE. The optimal parameter values for the synthetic database were chosen as the parameters with the smallest global RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (z_i - b_i)^2}{m}} \quad (2.10)$$

In equation (2.10),  $z_i$  refers to the estimated baseline and  $b_i$  to the simulated baseline.  $m$  is the signal length.

In order to compare the three algorithms, figure 2.7 shows a boxplot of the RMSE distribution for the different methods in their optimal settings. As we had already

seen in 2.6, while the *airPLS* method is able to improve the *ALS* approach, our *psalsa* performs better.

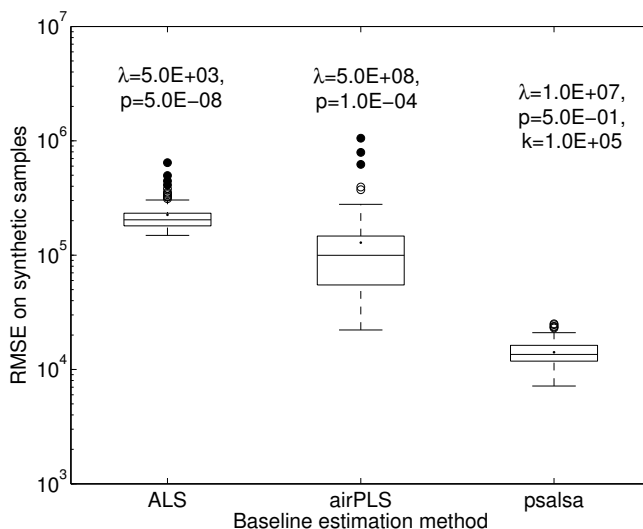


Figure 2.7: Comparison of the three methods for synthetic chromatograms

The main easy to argue weakness of the *psalsa* algorithm is the addition of a parameter that needs to be tuned. Figure 2.8 shows how the RMSE changes according to the value of  $k$  used. While we recover the performance of *ALS* and *airPLS* in the worst cases, in a range of three orders of magnitude the *psalsa* performance is better.

### 2.2.3.2 Real samples

Finally, we subjectively checked the baselines on urine samples from the real dataset. The three methods were applied to real samples. Figure 2.9 and figure 2.10 show the estimated baselines on different regions of a real urine sample. The first figure shows a region with large peaks. To avoid the effect present in *ALS* and shown in figure 2.5 where the baseline penetrated in the large peaks, a lower value of  $p$  had to be used. This is why the *ALS* baseline is always fitted below the noise level.

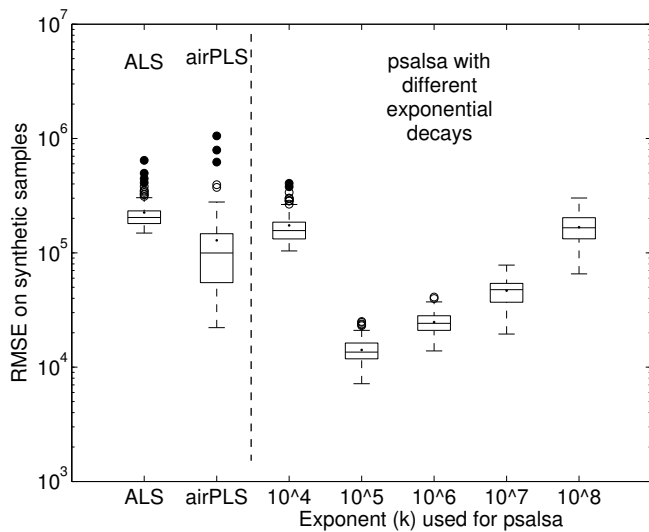


Figure 2.8: Performance comparison of different exponents for *psalsa*. ALS and airPLS optimal results are shown for comparison

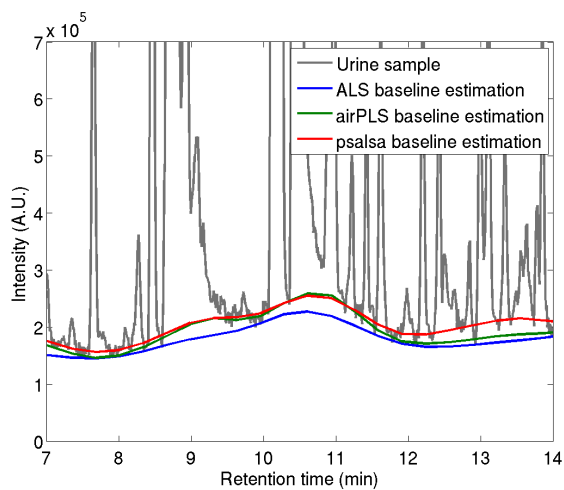


Figure 2.9: Comparison of the baseline corrections applied to real samples

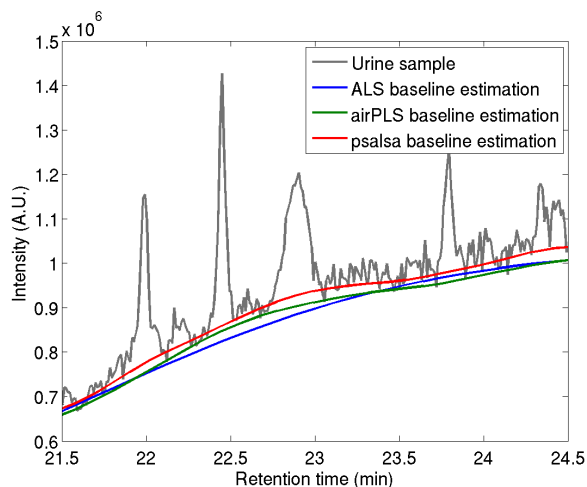


Figure 2.10: Detailed region of the comparison of the baseline corrections methods on real samples

## 2.2.4 Discussion and remarks

As mentioned, the original ALS algorithm was not designed specifically to fit signals with peaks several orders of magnitude above the baseline. Considering eq. (A.2), even though a small value for  $w_i$  is given for  $d_i > 0$ , given a large enough  $d_i$ , its contribution to  $S$  may still be dominant, producing an estimation of the baseline which contains part of the peak area. This forces us to choose a value for  $p$  so as the baseline does not penetrate in the peaks, instead of choosing  $p$  to cut through the noise as suggested in (Eilers Paul H. C., 2005). The value for  $p$  will then be smaller, leading to baseline estimations below the real baseline. Given that the estimation is below the baseline, a flexible baseline will be easier to adapt to the real baseline whenever possible, that is the reason why  $\lambda$  values are smaller in the ALS method with respect to the other methods.

Therefore, on the analysed signals, the parameters which minimize the RMSE on the ALS method are chosen to be able to properly fit the large peaks, instead of according to their theoretical purpose.

On the other hand, the *airPLS* algorithm is able to cope with large peaks, as it gives  $w_i = 0$  for  $d_i > 0$ . Unfortunately, that approach again leads to baselines fitted below the noise level instead of cutting through it. The *airPLS* algorithm was designed with the aim of removing the  $p$  parameter, and indeed  $p$  contribution



is less relevant to the final estimation than the contribution of  $p$  at the original ALS algorithm, as it is only used at the boundaries of the signal.

Finally, *psalsa* algorithm does not suffer the issues of the original *ALS* method, as the exponential modulation reduces the contribution to  $S$  of the large peaks. This makes it possible to use  $p$  to enforce that the baseline crosses the noise level, instead of fitting below it.  $p$  value is not comparable directly to the *ALS* method, as its contribution is modulated by the exponential. Even though *psalsa* requires an additional parameter ( $k$ ) to control the exponential decay of the weights, figure 2.8 shows that the RMSE value is smaller on *psalsa* on a range of three orders of magnitude, making it easy to provide a value for  $k$  that improves ALS results.

When applying the three methods on real samples, we can confirm how *psalsa* is able to estimate a baseline cutting through the noise, instead of being under-fitted as happens with the other two methods.

In chapter 4 (Applications), we will see how this method is applied to both GC-IMS samples and GC-MS samples. Here we will continue with the preprocessing techniques, discussing the alignment issues of GC-IMS samples.

## 2.3 Alignment

Spectral misalignments are a major concern of most spectroscopy and analytical chemistry instrumentation. Peaks of different samples corresponding to the same analyte should, in ideal conditions, appear in the same spectral position. However in many cases, we have to shift or warp the spectra to achieve that. This procedure is called alignment.

As mentioned in the introduction (see 1.4.1.3), there are two general strategies for spectral alignment: *Peak matching* and *Spectral warping*. Peak matching strategies are quite common in LC-MS and GC-MS data analysis and very common if tandem MS setups are used. The reason for choosing this strategy in those cases is that the samples are characterized by narrow sparse peaks, and therefore the approach of integrate peaks and then match them is much more computationally effective than looking for correlations in the whole spectra. Examples of peak matching strategies are found for instance in MzMine (Pluskal et al., 2010) where they use the RANdom SAMpler Consensus alignment technique, or PyMS (O'Callaghan et al., 2012) where they use a dynamic programming approach based on the Needleman-Wunsch algorithm (Robinson et al., 2007) for sequence alignment. In XCMS (Smith et al., 2006), the authors use a non linear method based on the clustering of peaks already identified by their mass spectrum.

On the other hand *spectral warping* is more often used in spectroscopy and in chromatography (either without further mass spectrometry fragmentation or applied to the total ion chromatogram). In metabolomics applications, Nuclear Magnetic Resonance (NMR) is one of the golden standard techniques where spectral warping techniques are common. A recent review by (Vu and Laukens, 2013) summarizes alignment methods applicable to NMR data and their main features. They describe 18 alignment methods, 6 of them using a peak picking based approach and 12 of them using spectral warping techniques. The spectral warping techniques can be classified according to the correction they apply (either shifting spectra, a polynomial correction, stretching, compressing...), the figure of merit or criteria used to determine the right alignment (Pearson correlation, FFT cross-correlation...) and whether or not the alignment technique is applied in segments or to the whole spectra at once. Among these techniques, *icoshift* (Savorani et al., 2010), (Tomasi et al., 2011), *Parametric Time Warping* (Bloemberg et al., 2010) and *Correlation Optimized Warping* (Nielsen et al., 1998) have gained popularity thanks to their simplicity and availability of their implementations.

Even though many methods exist, conventional data acquisition software for GC-IMS data processing (e.g. LAV<sup>4</sup>, from GAS Dortmund) often provides simple and/or manual alignment methods, sometimes limited to linear distortions of the retention time and drift time axes. With these methods it is not possible to fully align the spectra, as for instance retention time variations are typically non linear.

Misalignments in retention time are produced by instrumental drift related to column degradation, as well as variations in temperature, pressure or flow. Each of these misalignments is characterized by a different time scale. Column degradation affects the measurements on a monthly time scale, while flow variations can happen between measurements or during a measurement.

In this section we characterize the misalignments of GC-IMS samples in a more than 10 months long study, using external calibrants. This characterization shows the feasibility of the sample alignment even with samples measured several months apart, and allows us to discuss how often is it worth analyzing external calibrant samples in a GC-IMS study. Based on the results, we discuss the usefulness of measuring external calibrant samples and its limitations. To perform the retention time alignment, both a linear method and a non-linear method are used to highlight the advantage of the non-linear method, based on monotonic cubic splines.

---

<sup>4</sup><http://www.gas-dortmund.de/index-gas.php?lan=1&spath=463>

### 2.3.1 Dataset

For the alignment analysis presented in this section, we used calibrant samples. These samples are a mixture of six ketones (2-butanone, 2-pentanone, 2-hexanone, 2-heptanone, 2-octanone, and 2-nonanone), measured in 15 days. These 15 days span in an irregular pattern across 10 months. The experimental distribution allows us to explore both the short term (same-day, few days apart) and long term (weeks, months) misalignments.

The samples were analyzed by Dr. Lourdes Arce’s research group at the Department of Analytical Chemistry from the Universidad de Córdoba. The analytical protocol summarized here is explained at (Garrido-Delgado et al., 2015). The GC-IMS instrument is a commercial FlavourSpec® model from Gesellschaft für Analytische Sensorysysteme mbH (G.A.S., Dortmund, Germany) with a 30 m long  $\times$  0.25 mm (inner diameter) chromatographic column filled with 0.5  $\mu$ m film thickness of methyl, phenyl and vinylsiloxane from CS-Chromatographie Service GmbH (Düren, Germany) in a 94 : 5 : 1 proportion. The injection rate was 100  $\mu$ L/s, and the carrier flow rate was set to 5 mL/s. The column was operated under isothermal conditions at 40 °C. The spectrometer was equipped with a heated splitless injector with 2 mm inner diameter, 6.5 mm outer diameter  $\times$  78.5 mm fused quartz glass. This enabled direct sampling of the headspace from the samples by using a 2.5 mL Hamilton syringe furnished with a 51 mm needle of 23 gauge from CTC Analytics AG (Zwingen, Switzerland). The inlet septa injector used was 11 mm in diameter and supplied by Agilent Technologies (Santa Clara, CA, United States). The instrument was also coupled to an autosampler unit from CTC Analytics AG (Zwingen, Switzerland).

For analysis, the ketones mixture was prepared in a 20 mL vial that was closed with magnetic caps. After 8 min of incubation at 60 °C, 200  $\mu$ L of sample headspace was automatically injected by means of a heated syringe (80 °C) into the heated injector (80 °C) of the GC-IMS equipment. After injection, the nitrogen gas used as carrier gas, with inlet pressure of 4 bars, passed through the injector inserting the sample into the gas column, which was heated at 40 °C for timely separation. Then, the analytes were eluted in the isothermal mode and driven into the ionization chamber for ionization, prior to spectrometric detection. Molecules were ionized using a Tritium source (6.5 keV) and the resulting ions driven to the drift region via a shutter grid (Bradbury and Nielson design), set at a pulse width of 100  $\mu$ s. The drift tube was 5 cm long and operated at a constant voltage of 400 V  $\text{cm}^{-1}$ , a temperature of 45 °C, and a drift gas flow rate of 250 mL  $\text{min}^{-1}$  (Nitrogen). Data were acquired in the positive ion mode. The detector offered a sampling frequency of 150 kHz in the drift time axis. A

full ion mobility spectrum was acquired every 21 ms, and the firmware was set to record the average of 32 spectra for noise reduction purposes, leading to a sampling period in the retention time axis of 672 ms.

Table 2.1: Time distribution of 44 calibrant samples

Date	Time since previous	Number of Samples
2015-03-16		1
2015-03-20	4 days	1
2015-03-23	3 days	1
2015-03-31	8 days	1
2015-04-06	6 days	1
2015-04-14	8 days	1
<b>2015-05-06</b>	<b>&gt;3 weeks</b>	<b>1</b>
2015-05-14	8 days	1
<b>2015-12-03</b>	<b>&gt;6 months</b>	<b>5</b>
2015-12-09	6 days	5
2015-12-10	1 days	5
2015-12-14	4 days	4
2015-12-16	2 days	5
<b>2016-01-11</b>	<b>&gt;3 weeks</b>	<b>7</b>
2016-01-14	3 days	5

During the time of the analysis, the instrument was operated regularly, analyzing the headspace of olive oil samples so conventional instrument degradation and fluctuations are to be expected. Table 2.1 shows the distribution of the calibrant samples, as well as the gap with respect to the previous calibrant analysis.

### 2.3.2 Methodology

The data analysis strategy is represented in figure 2.11. As explained in previous sections, samples were denoised using a second order Savitzky-Golay filter with a 19 point window (0.12 ms), applied to each IMS spectra, and the baseline was removed using the presented *psalsa* algorithm ( $\lambda = 10^6$ ,  $p = 0.005$ ). Peaks were easily detected based on the position of the maximum in the regions where each of the ketones were expected to appear. Some of the ketones presented two ion clusters, as a monomer and a dimer and in those cases they are reported independently. At this point, a table with (sample, ketone, cluster, retention

time, drift time) was built, where all the retention times and drift times for a given peak were expected to be the same under ideal conditions. The warping functions to align the spectra in both axis were based on the alignment of these ketone peak positions.

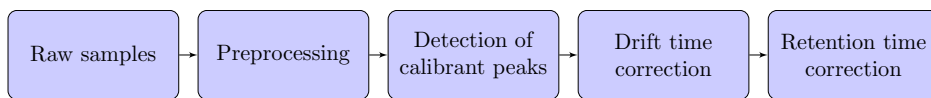


Figure 2.11: Data analysis flow chart for the alignment study

**Drift time:** The drift times were compensated using a multiplicative correction. The multiplicative correction is a simple and linear approach to correct the drift time. It assumes a warping function of the drift time based on a linear relation  $t' = kt$ , where the  $k$  has to be estimated. The estimation of the right  $k$  in the linear transformation was based on the position of the Reactant Ion Peak, so all the reactant ion peaks were properly aligned. This correction is compatible with the conversion from drift times to reduced mobilities explained on section 1.3.1.1. From equations (1.1) and (1.2), we can derive the reduced mobility as:

$$K_0 = \frac{L}{Et_d} \frac{273}{T} \frac{P}{760} = \frac{273LP}{760ET} \frac{1}{t_d} \quad (2.11)$$

Equation (2.11) shows how the conversion from drift time to reduced mobility is mathematically equivalent to a linear transformation of the drift time axis.

**Retention time:** To align the retention times, a monotonic cubic spline interpolation method was used. To the best of our knowledge, the use of monotonic cubic splines for the alignment of GC-IMS data has not been used in the past, however in chromatography applications the most similar work we were able to find was reported by (Halang et al., 1978), where natural cubic splines are used for the alignment of retention indices. More recently (Eilers, 2004), highlight the non-uniform retention time distortions reported at (Gong et al., 2004) on a High Pressure Liquid Chromatography - Diode Array Detection (HPLC-DAD) instrument, and suggest as a possible alignment strategy the use of p-splines (penalized B-splines).

Given that we know the retention time of each of the ketones for each sample, it would be desirable that the warping function is flexible enough to fit those variations. Also, the elution order is not expected to change (as cross-over retention time effects (Mehran et al., 1991) are not expected in these calibrants), so it is feasible to impose in this case the monotonic assumption. Finally, changes to the retention time must be “smooth” as rough transitions would severely distort our

spectra and “local” meaning that the warping corrections done on a retention time region should be mainly influenced by points in the vicinity of that region and not by calibrants of analytes eluting at distant retention times.

Linear or polynomial models do not have enough flexibility to fit the retention time variations, and they have limited locality. The locality can be increased by using piecewise linear models or piecewise polynomial models, but piecewise warping functions are not necessarily differentiable at the edges. This is a general problem of all the alignment algorithms based on segments: On the segment boundaries artifacts occur due to the non differentiability of the warping function. For instance, the popular `icoshift` algorithm (Tomasi et al., 2011) resorts to providing missing values on segment edges to prevent the appearance of artifacts. Monotonic cubic splines as provided by (Hyman, 1983) fulfill all the conditions. They are continuous and differentiable and at the same time they are able to fit the calibrant points that we use as support. The `splinefun` function of the `stats` R package, provides an easy to use implementation as well.

### 2.3.3 Results and Discussion

After peak detection, it is easy to visualize the position of the unaligned peaks of the calibrant samples. Figure 2.12 shows how the most important misalignments appear in the retention time axis, while the drift time axis has smaller fluctuations. The Reactant Ion Peaks are shown as a straight vertical line around  $6.5\text{ ms}$  and present a dispersion comparable to the ketone peaks. The mean position of the ketone peaks as well as their dispersion, represented with the standard deviation is shown on table 2.2.

Table 2.2: Raw peak positions

Name	Cluster	Drift time (ms)	Retention Time (s)
2-butanone	Dimer	$8.58 \pm 0.11$	$92 \pm 11$
2-pentanone	Dimer	$9.44 \pm 0.11$	$124 \pm 15$
2-hexanone	Dimer	$10.35 \pm 0.12$	$212 \pm 27$
2-heptanone	Dimer	$11.24 \pm 0.13$	$401 \pm 33$
2-heptanone	Monomer	$8.66 \pm 0.11$	$406 \pm 32$
2-octanone	Dimer	$12.12 \pm 0.13$	$584 \pm 36$
2-octanone	Monomer	$9.15 \pm 0.11$	$590 \pm 36$
2-nonanone	Dimer	$12.96 \pm 0.13$	$1052 \pm 88$
2-nonanone	Monomer	$9.65 \pm 0.11$	$1053 \pm 87$

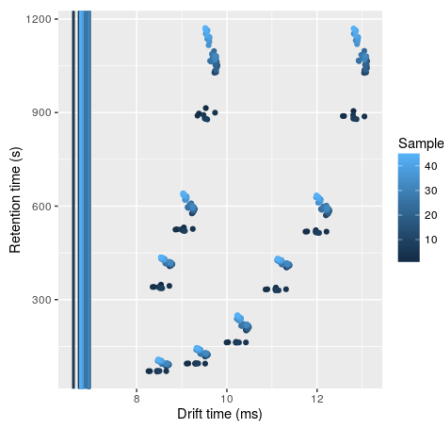


Figure 2.12: Peak positions for the 44 calibrants, both in retention time and drift time. On the left, there is a band of misaligned RIPs.

As expected, smaller ketones had shorter drift times, as they can travel faster through the drift tube. Similarly they also elute faster from the chromatographic column. With respect to the monomer/dimer differences, apart from the obvious shorter drift times in monomer ions, in some cases the monomer was detected a bit earlier than the dimer. This is the normal and expected behavior described at section 1.3.1.2.

Considering instrumental drift, the retention time axis shows a long term drift behavior. This can be seen at figure 2.12, where there is a clear drift of each cluster of points towards larger retention times that is correlated with the day of the analysis. This drift also exists in the drift time axis, albeit with a major random contribution.

### 2.3.3.1 Drift time correction

Figure 2.13 represents the peak positions after the drift time alignment. The RIP of all the samples is overlapped, as it is the reference peak used for the estimation of  $k$  in the linear distortion. The drift time dispersion is greatly reduced. Table 2.3 shows how the standard deviation of all the clusters is reduced by almost an order of magnitude in the drift time axis after the correction.

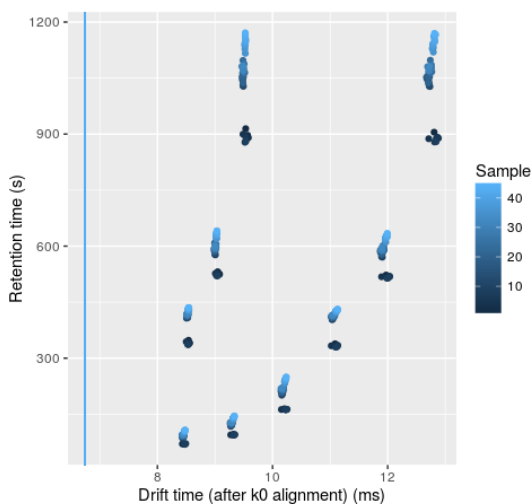


Figure 2.13: Peak positions for the 44 calibrants, both in retention time and drift time. The drift time axis is aligned, and RIPs match.

Table 2.3: Peak positions after drift time correction

Name	Cluster	Drift time (ms)	Drift time (corr.) (ms)
2-butanone	Dimer	$8.58 \pm 0.11$	$8.46 \pm 0.02$
2-pentanone	Dimer	$9.44 \pm 0.11$	$9.3 \pm 0.02$
2-hexanone	Dimer	$10.35 \pm 0.12$	$10.19 \pm 0.03$
2-heptanone	Dimer	$11.24 \pm 0.13$	$11.07 \pm 0.04$
2-heptanone	Monomer	$8.66 \pm 0.11$	$8.53 \pm 0.01$
2-octanone	Dimer	$12.12 \pm 0.13$	$11.94 \pm 0.05$
2-octanone	Monomer	$9.15 \pm 0.11$	$9.02 \pm 0.02$
2-nonanone	Dimer	$12.96 \pm 0.13$	$12.76 \pm 0.05$
2-nonanone	Monomer	$9.65 \pm 0.11$	$9.51 \pm 0.03$

Finally, figure 2.14 shows the estimated correction factor  $k$  for each of the analysis, sorted and colored by day. The correction factor is between 0.97 and 1.03 in all cases, meaning that the distortion is below  $\pm 3\%$  under normal working conditions. It is worth mentioning that samples analyzed on the same day tend to cluster together in the correction factor estimation indicating that drift time misalignments in the same day are smaller than in different days.



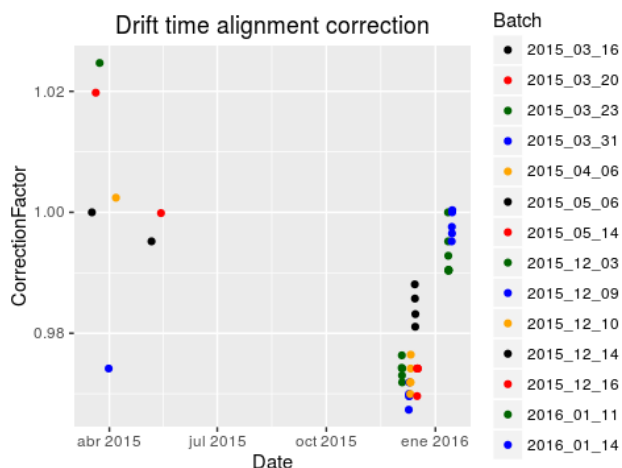


Figure 2.14: Correction factor for the drift time alignment

### 2.3.3.2 Retention time correction

After the alignment on drift time it is time we proceed to align the retention times. We will compare a simple linear regression warping with the monotonic cubic splines. The linear regression as a basis for the comparison is chosen for its availability in commercial GC-IMS data analysis software and because of its simplicity. If the linear regression warping provides a first order alignment correction, then the monotonic cubic splines should improve the alignment thanks to the extra flexibility that the splines can give, but in any case should be a second order improvement to the linear regression. Having these two methods makes it easy to understand where and how monotonic cubic splines improve the linear regression warping.

One of the concerns of using splines (an interpolation method) instead of fitting a linear model is the risk of overfitting. With great *freedom* comes great responsibility, and having more degrees of freedom to adjust the splines requires the responsibility to validate the alignment in a fair way. To this end, the following procedure was chosen to ensure a fair validation of the algorithm:

- The oldest sample was chosen as a reference
- A warping function  $w(t)$  was estimated for the first sample of each batch with respect to the chosen reference
- The warping function was applied to the rest of the samples in the batch.

- **Validation:** The procedure was repeated excluding one of the ketones from the analysis and predicting its position.

The first result of applying the methodology using a linear model and using monotonic cubic splines is the comparison of the appearance of the warping functions. Figure 2.15 shows on the left the linear warping functions for all the samples while on the right there is the monotonic cubic splines interpolation solution. While both methods provide similar warping functions, it can be seen how the monotonic cubic splines have more flexibility and provide second order corrections to the alignment, especially around  $t_r = 200$  s and around  $t_r = 750$  s.

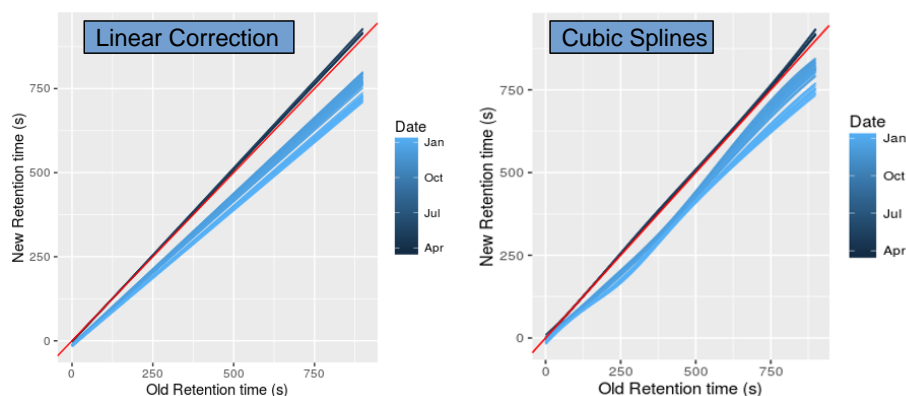


Figure 2.15: Warping functions for linear and cubic retention time corrections

The next figure to consider is what happens to the retention time estimations if we leave a ketone out. By leaving a ketone out we are in practical terms trying to align a peak that is far from the calibrants. Table 2.4 shows the comparison of the estimation of the excluded ketones. Note that the 2-butanone and 2-nonanone were not left out because they were the calibrant extremes of the analysis. This table shows the limitation of the linear model when trying to fit any of the peaks. The average predicted retention times for the linear correction method presents a much larger bias than the monotonic cubic splines method. This is more clear in the case of the 2-octanone. One explanation for this is that the 2-octanone is a point with a lot of leverage in the linear model and therefore the linear model suffers more to predict its position when it is missing. On the other hand, the flexibility of the splines model allows to have less bias in the prediction of the ketones position. For the monotonic cubic splines model, the dispersion is similar to the linear model, because of the leverage of the 2-octanone peaks.

Table 2.4: Comparison of the estimated retention times

Peak left out	Reference (s)	Linear (s)	Splines (s)
2-pentanone	95	$98 \pm 4$	$94 \pm 3$
2-hexanone	165	$178 \pm 8$	$169 \pm 5$
2-heptanone	343	$345 \pm 10$	$342 \pm 11$
2-octanone	527	$502 \pm 16$	$531 \pm 21$

One way to visualize the effect of the retention time alignment is through the reverse RIP, described in section 1.3.2.2. Figure 2.16 shows the reverse RIP for each of the samples. On the left column, a large retention time region is shown, while a more detailed retention time region is represented on the right. From top to bottom, there are the three described scenarios: before the correction, using a linear model and using monotonic cubic splines. The advantage of the monotonic cubic splines is remarkable, especially for larger retention times.

Having assessed the performance of the monotonic cubic splines and compared it to a linear model, the final question to be answered is if we can make a recommendation for a recommended time lapse between the analysis of an external calibrant. The answer to this questions boils down to answering “For how long can we use the same calibrant in order to align the future samples to the same reference?”. If the answer to this question is “hours”, then it may be useful to sample a calibrant at the beginning and at the end of each analysis day. On the other hand, if the answer is “months”, then doing several calibrant analysis per day is simply a waste of time.

To answer this question we used our dataset of calibrant samples, we computed a warping function between a reference sample and a mapping sample and we applied the warping function to future samples measured that same day, the next day, etc. By measuring the error in the prediction of the retention times, we are able to represent in figure 2.17 the relative error in the retention time estimation with respect to the number of days elapsed between the date when the sample that was used to calculate the warping function was measured and the date when the corrected spectra was measured.

The figure shows a bias starting from the second day, indicating that measuring one calibrant every one or two days is enough to be able to compensate for the instrumental drift. Having more than one calibrant sample per day is not very useful, as the random fluctuations between consecutive samples are larger than the intra-day variability. On the other hand, if we only take one calibrant sample at the beginning of each month we won’t be able to fully correct the retention time drifts by using calibrants.

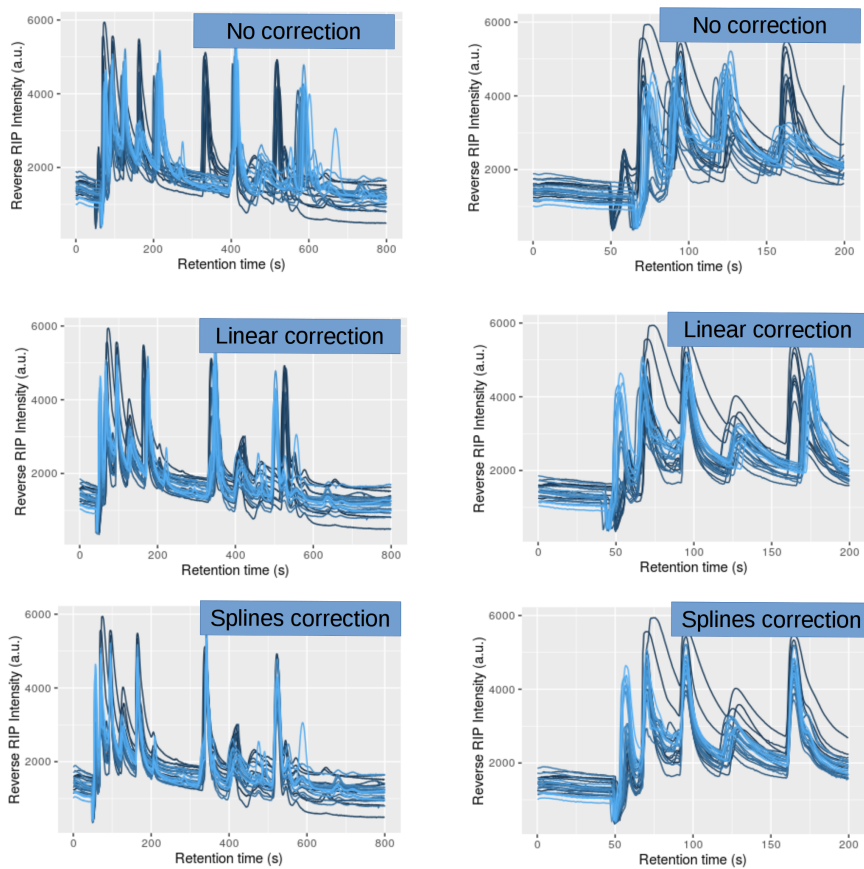


Figure 2.16: Reverse RIP alignment result across multiple samples.

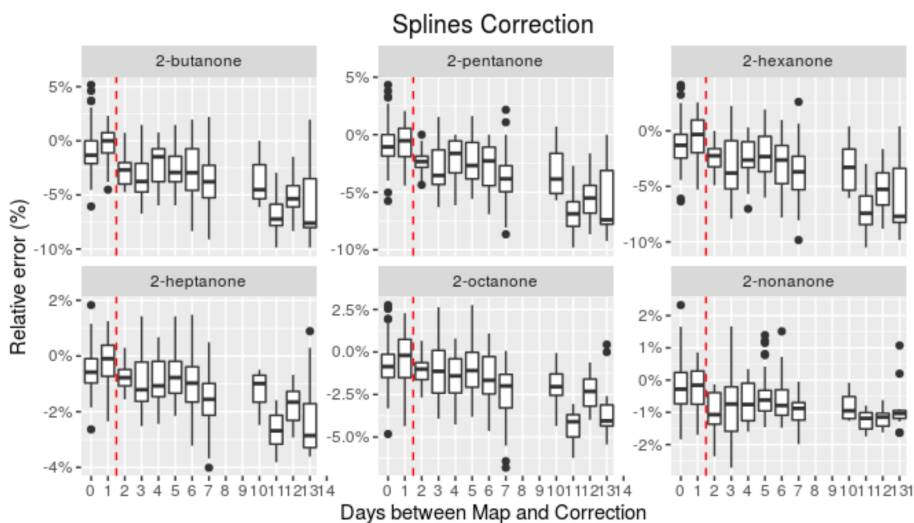


Figure 2.17: Relative prediction error depending on how old is the calibrant sample used to train the model.

This analysis has contributed to characterize both the drift time and the retention time misalignments. The use of monotonic cubic splines for the retention time alignment of GC-IMS samples has been presented, showing that it offers a simple improvement over the often used spectral shifts and linear corrections. We also suggested a reasonable time span of two days as the optimal period for the analysis of two consecutive external calibrant samples. A clear limitation of this study is the lack of exploration of internal calibrants for sample alignment. The use of internal calibrants makes the alignment problem easier, as all samples share the same set of calibrant analytes. However, the use of internal calibrants is often more complex, as there is the need for choosing a calibrant that does not interfere with the sample under study or its matrix. While this is much easier on targeted studies (where peaks are expected at some specific positions, and calibrants can be selected to avoid interference with those positions), on untargeted studies those “free” regions may not be as easily defined, especially on the ion mobility axis, that presents wider peaks and non-linear effects with mixtures due to charge competition.

## Chapter 3

# Sliding Window Multivariate Curve Resolution for GC-IMS data

Blind Source Separation (BSS) techniques aim to extract a set of source signals from a measured mixture in an unsupervised manner. In the chemical instrumentation domain, source signals typically refer to time-varying analyte concentrations, while the measured mixture is the set of observed spectra. Several techniques exist to perform BSS on Ion Mobility Spectrometry, being Simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) and Multivariate Curve Resolution (MCR) the most commonly used. The addition of a multi-capillary gas chromatography column using the ion mobility spectrometer as detector has been proposed in the past to increase chemical resolution. Short chromatography times lead to high levels of co-elution, and ion mobility spectra are key to resolve them. In this chapter, BSS techniques are used to deconvolve samples of the gas chromatography - ion mobility spectrometry tandem. We propose a method to extract spectra and concentration profiles based on the application of MCR in a sliding window. Our results provide clear concentration profiles and pure spectra, resolving peaks that were not detected by the conventional use of MCR. The proposed technique could also be applied to other hyphenated instruments with similar strong co-elutions.

## 3.1 Introduction

BSS techniques, also named in chemometrics “resolution techniques”, are commonly applied to hyphenated analytical techniques that provide second order data. In IMS samples, the compounds’ original concentration profiles and pure spectra can be deconvolved from the sample using BSS techniques (Pomareda et al., 2010). For the first time, we propose a blind source separation technique in MCC-IMS data. Direct application of MCR techniques to full MCC-IMS data typically fails to resolve co-elution due to the complexity of the data and to the global noise which hinders the detection of weak but significant peaks. The typical approach in this case is the manual selection of the retention time window where the co-elution appears and the application of MCR in this data subset. However, in MCC-IMS chromatography conditions, co-elution is a prevalent phenomenon (Eiceman et al., 1995), (Baumbach, 2009). Few individual peaks can be isolated in the total chromatogram and mostly very broad peaks are observed. To deal with this complexity, we propose an automatic manner to investigate co-elution across the whole chromatographic axis. The proposed method is able to detect and recover compounds in adverse co-elution conditions and reject spurious spectra with no physical meaning in an unsupervised manner.

The method is applied to real data corresponding to olive oil headspace analysis, with the aim to extract accurate concentration profiles and pure spectra for each sample. The extracted information can be used later on to discriminate among different regulated olive oil qualities in fraud prevention applications.

### 3.1.1 Blind Source Separation techniques

Blind source/signal separation techniques are the collection of algorithms designed to estimate a set of source signals from measured mixtures. As mentioned in (Cardoso, 1998), techniques are blind because a) the source signals are not observed directly, b) the mixing matrix is unknown and c) no information is available about the composition of the mixture, not even the number of source signals present. These techniques are commonly used in signal processing (Cichocki and Amari, 2002) and are increasingly being used in chemical instrumentation applications (Duarte et al., 2014), such as the analysis of nuclear magnetic resonance data (Nuzillard and Nuzillard, 1998), chemical reaction monitoring (Carteret et al., 2009) and Raman spectroscopy (Miron et al., 2011). BSS has been recently used to enhance information extraction from temperature-modulated metal oxide gas sensors (Montoliu et al., 2010) and to separate interferences from ion activity in ion-sensitive field-effect transistors (Bermejo et al., 2006).

As deconvolution problems are under-determined by definition, constraints are required to narrow the space of solutions. For many applications, Independent Component Analysis (ICA) (Hyvärinen et al., 2001) is an appropriate and successful technique if mixing models can be assumed to be linear and source signals to be statistically independent. However, in chemical analysis and specifically in IMS, statistical independence of compounds is not necessarily fulfilled (Pomareda et al., 2010). Therefore other approaches are used to constrain the range of possible solutions (Duarte et al., 2014) being Non-negative Matrix Factorization (NMF) techniques (Cichocki et al., 2006) and in particular Multivariate Curve Resolution (MCR) methods (Lawton and Sylvestre, 1971) common alternatives.

In MCC-IMS applications, BSS techniques are helpful when several analytes elute at the same time from the MCC and they are detected by the IMS. If some of the co-eluting analytes present larger proton affinities (or electronegativities), they can mask and hide the rest of the analytes due to the charge competition effect described on section 1.3.1.2. In this case, no posterior data analysis technique (BSS or other that we know of) will be able to detect them.

### 3.1.2 Multivariate Curve Resolution Alternating Least Squares

Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) (Tauler, 1995) assumes a linear decomposition of the mixing matrix, which can be written as shown in Eq. (3.1).

$$D = CS^T + E \quad (3.1)$$

- $D$  ( $M \times N$ ) is the measured mixing matrix, with  $M$  spectra of length  $N$ .
- $C$  ( $M \times K$ ) is the abundances or concentrations matrix, that contains the proportions of each unmixed spectrum in the measured matrix and,
- $S$  ( $N \times K$ ) is the pure (or unmixed) spectra matrix that contains the  $K$  pure spectra of length  $N$ .  $E$  is a matrix of residuals ( $M \times N$ ).

Given an initial estimation of  $K$  pure spectra, MCR-ALS proceeds as follows:

1. Filter noise from the mixing matrix: First, compute PCA scores and loadings from the  $D$  mixing matrix. Then reconstruct a filtered version of  $D$ , named  $D^*$ , using the first  $K$  principal components of the computed scores and loadings.
2. Estimate the concentration profiles using least squares:



$$C = \operatorname{argmin}_C \|D * -CS^T\|^2 \quad (3.2)$$

3. Impose constraints on the concentration profiles
4. Estimate the pure spectra using least squares:

$$S = \operatorname{argmin}_S \|D * -CS^T\|^2 \quad (3.3)$$

5. Impose constraints on the pure spectra
6. Iterate steps 2-5 until convergence.

The key to obtaining reliable concentration profiles and pure spectra depends on the estimation of the number of components in the mixture, the initialization of the pure spectra and the imposition of constraints.

The number of components for each window can be estimated with several methods such as (Buxton and Harrington, 2001), (Windig et al., 2005), (Gourvénc et al., 2002). However, a simpler approach described in (Diewok et al., 2003) is commonly used: the number of components is determined as the number of singular values of the matrix above a given threshold, representative of the noise in the sample.

There are multiple ways of obtaining an initial estimation of the pure spectra, being SIMPLE-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) (Windig and Guilment, 1991) and Evolving Factor Analysis (EFA) (Maeder, 1987) the most common ones. Although EFA works well on samples presenting unimodal concentration profiles, on IMS this condition does not necessarily hold, making SIMPLISMA the most common alternative in this field (Pomareda et al., 2010), (Harrington et al., 1997).

Many constraints can be imposed on the concentration profiles and pure spectra depending on the prior knowledge of our particular problem: On IMS spectra, non-negativity can be imposed on both concentration profiles and spectra. Moreover, as the ionization process consists in a charge transfer from the RIP to the compounds, charge conservation can be imposed on the concentration profiles (closure on C). Unimodality constraints are not suitable for concentration profiles, but can be imposed to the resolved spectra shapes. Finally, selectivity constraints to the concentration profiles can also be imposed if some components are known to appear at a particular retention time range.

The MCR-ALS algorithm is based on a least squares minimization of the global error of the factorization. As it is shown in our results, local peaks with low intensities appearing in regions with strong co-elution may pass unnoticed by

MCR-ALS, as they present a contribution to the error comparable to or smaller than the global noise of the sample. In these cases, increasing the estimated number of components in the mixture leads to extracting spurious compounds with no physical meaning instead of the desired local compounds.

### 3.1.3 Proposed technique: Sliding Window Multivariate-Curve Resolution

In order to overcome the limitation in resolving low intensity peaks when the conventional MCR-ALS is applied to the whole MCC-IMS data matrix, we propose to apply MCR in short partially overlapped windows, slicing the matrix in the retention time axis. In addition, window overlap is imposed to avoid splitting peaks on window borders and to avoid detecting spurious compounds inconsistent across windows.

First, the initial estimations of pure spectra and concentration profiles are obtained by applying SIMPLISMA to each window. By using SIMPLISMA in this fashion, we can extract local peaks with low intensities, as they have comparatively higher peak purity within a single window. The number of components for each window is estimated using the threshold on singular values previously described in Section 3.1.2. To select the threshold, the singular values were plotted in decreasing order (plot not shown), presenting the typical elbow-like shape. The threshold was selected when the singular values begin to stabilize. Given the initial estimations, we use MCR-ALS to extract a set of concentration profiles and pure spectra for each window.

Finally, the results from all the windows are merged into a single set of concentration profiles and spectra representative of the whole sample. To do so, compounds are tracked through consecutive windows based on the similarity of their spectra. The angle between two pure spectra  $s_i$  and  $s_j$  is computed as shown in Eq. (3.4).

$$\theta_{i,j} = \arccos \left( \frac{s_i s_j}{\|s_i\| \|s_j\|} \right) \quad (3.4)$$

Figure 3.1 shows a diagram with an example of four compounds being tracked along three windows. The link between two spectra of consecutive windows is formed only if their angle is below a given threshold. In this figure, compounds C1 and C2 are being tracked along all the windows (N to N+2) while compound C3 disappears on window N+1 because no link can be established on window N+2. Compound C4 does not appear until the N+1 window. The last spectrum

in window N+1 does not establish any link, thus it is considered spurious and is rejected from the final set of tracked compounds.

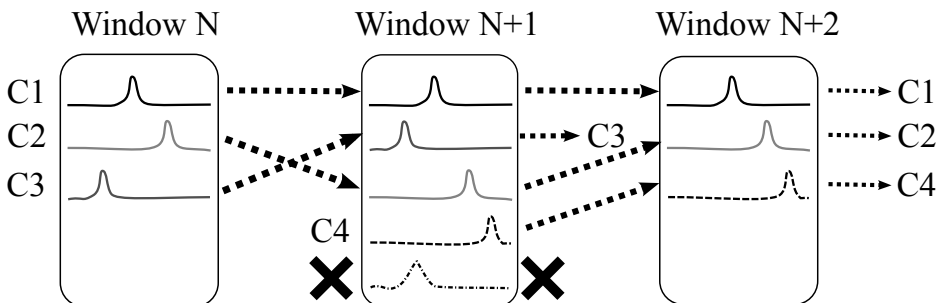


Figure 3.1: Diagram of the tracking of spectra through three windows. Links between spectra are established if their angle is lower than a given threshold.

Windows are highly overlapped to guarantee that if no link can be established for a spectrum, then it can be safely considered as spurious and rejected from the final set. The final estimation of the pure spectra for each compound is computed as the mean of all the tracked spectra. The standard deviation of the mean is used as its error estimation. Averaging and computing the standard deviation are used likewise to obtain the final estimation of the concentration profiles.

## 3.2 Materials and Methods

### 3.2.1 Description of the samples

The proposed technique was applied to the olive oil dataset described in (Garrido-Delgado et al., 2012). Current regulations in the European Union classify olive oils in three different categories according to their quality, namely Extra Virgin Olive Oil (EVOO), Virgin Olive Oil (VOO) and Lampante Olive Oil (LOO), being EVOO the category of highest quality and LOO the lowest one. This classification is based on several chemical parameters (free acidity, peroxide value and Ultra-violet absorbance) and a sensorial analysis. A proper control of olive oil qualities is crucial, not only because of the difference in price but also because LOO is not suitable for human consumption without being previously refined.

Ninety-eight olive oil samples from different qualities (27 samples of LOO, 28 samples of VOO and 43 samples of EVOO) were obtained from the Agrarian Laboratory of Junta de Andalucía and an oil press from Córdoba (Spain) during

the 2009-2010 and 2010-2011 harvests. In order to keep the organoleptic features of the samples, they were stored at 4°C until their analysis.

### 3.2.2 Analytical methods

Samples were analyzed with a MCC-IMS instrument (FlavourSpec®) from Gesellschaft für Analytische Sensorysysteme mbH (G.A.S), Dortmund (Germany). The olive oil headspace was directly sampled with a heated splitless injector, and the instrument was coupled to an automatic sampler unit (CTC-PAL, CTC Analytics AG, Zwingen, Switzerland) to improve reproducibility.

One gram of sample was placed in a 20-mL vial that was closed with magnetic caps. Samples were incubated at 60°C for 10 minutes and 100  $\mu$ L of sample headspace was automatically injected into the injector (80°C) of the MCC-IMS.

The carrier gas going through the injector inserted the sample into the chromatograph, previously heated to 30°C for pre-separation on a non-polar OV-5 MCC (20 cm long, ~1000 parallel glass capillaries, filled with 5% diphenyl and 95% dimethylpolysiloxane). The analytes were eluted in an isothermal mode and driven into the IMS.

Inside the IMS, the ionization was produced with a Tritium source (6.5 keV). Ions entered the 6 cm long drift tube operating at a constant electric field of 350 V/cm and at a temperature of 60°C. Spectra were acquired in the positive ion mode, generating each spectrum with the average of 32 scans, using a grid pulse width of 100 ns. The IMS sampled at 150 kHz and each scan lasted 20 ms. Each spectrum is 3000 points long.

Each sample was analyzed for 15 minutes, obtaining a complete IMS spectrum every 0.7 seconds. Compounds only eluted during the first 4 minutes of the retention time, leading to 340 spectra with information per sample. Each sample can be represented by a 340x3000 matrix.

### 3.2.3 Pre-processing

Noise present in each spectrum of the sample was filtered using a second order Savitzky-Golay filter (Savitzky and Golay, 1964) with a window size of 13 data points. The window size was selected assessing that the RIP height distortion caused by the filter was smaller than 1% of its non-filtered maximum value.

Next, a baseline was estimated and subtracted from the spectrum: the estimation of the baseline was computed by fitting a 4th order polynomial to two non-peaked

(empty) regions in the spectrum found in the regions 1-5 ms and 14.7-18.7 ms.

Finally, only the drift time region from 4 ms to 14.65 ms (1600 sampled points) contained information, so irrelevant regions were cropped out. Each sample was therefore reduced to a 340x1600 matrix.

### 3.2.4 Sliding Window MCR

The proposed Sliding Window MCR (SW-MCR) technique is applied to the sample, using a window length of ten spectra (7 seconds) and a window shift of a single spectrum (0.7 seconds). The window length was selected based on the typical width of a peak in the chromatogram, computed as the median full width at half maximum (FWHM) of ten representative peaks in the sample.

Larger window sizes and smaller window overlaps may be used to reduce the computational cost of the method. A larger window size would imply that more compounds can be found in the same window. If the window is too large we will face the same problem than with conventional MCR-ALS application: we may fail to detect local peaks with low intensities. Regarding the window overlap, if the window overlap is too small this would increase our chances of splitting peaks in window borders and would hinder our ability to distinguish spurious solutions from actual compounds, as actual compounds would not have to necessarily appear among consecutive windows anymore.

After inspecting the distribution of singular values along the windows, we set a threshold to determine the number of components. Data not represented by the selected components was discarded using a PCA filter.

Regarding the MCR-ALS configuration, we initialized the pure spectra and the concentration profiles for each window using SIMPLISMA. We imposed the following constraints: 1) non-negativity to both concentration profiles and pure spectra via fast non-negative least squares, 2) closure to the concentration profiles and 3) unimodality to the resolved spectra. Additionally, we imposed a selectivity constraint to improve the RIP pure spectrum estimation: given that at the end of the sample (high retention times) no compounds elute from the column, the only compound present in the latest spectra is the RIP. From a blind source separation perspective, this information is very valuable, as an accurate estimate of the RIP pure spectrum can be easily obtained.

Finally, in order to track the resolved spectra through the windows, an angle threshold of 15 degrees was used. This angle was chosen after inspecting the angle distribution of the pairwise comparison of the spectra.

### 3.3 Results and Discussion

A selected informative region of an olive oil sample can be seen in figure 3.2. The Reactant Ion Peak can be seen close to 6 ms in drift time along all the retention time range. There are multiple peaks in the same retention time range, indicating a strong co-elution of the components. As peaks are created by the transference charge from the reactant ions, the RIP intensity decreases at the retention time when other peaks appear in the ion mobility spectrum. At higher retention times the RIP recovers all the charge returning to a constant intensity as no more compounds elute.

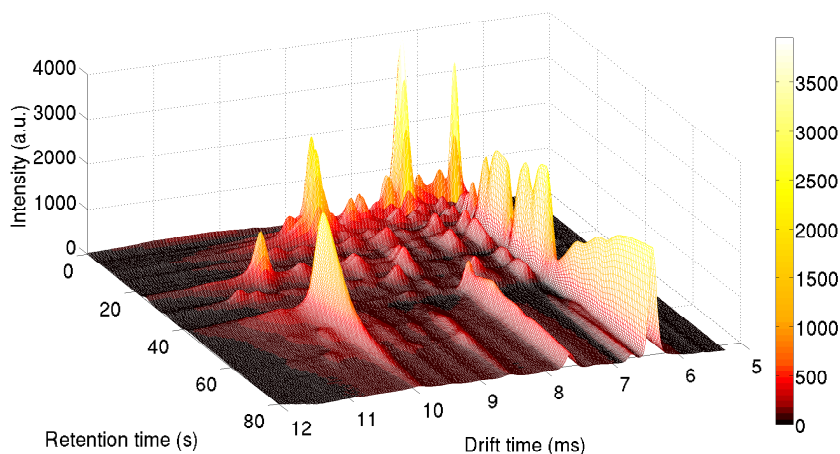


Figure 3.2: Region of a MCC-IMS olive oil sample. The RIP is observed at 6 ms. Multiple peaks on the same retention time indicate a strong co-elution. Note that both axes are reversed to prevent high intensity peaks from hiding the low intensity ones.

The intensity of the RIP can be used as a non-selective measure of the global elution of compounds. Integrating the RIP (from 6.26 ms to 6.6 ms) and subtracting it from the maximum intensity, we obtain the charge that has been transferred to other compounds throughout the retention time. This figure of merit is called the “Reverse RIP” and it is analogous to a total ion chromatogram in gas chromatography - mass spectrometry samples. Fig. 3 shows the reverse RIP of an olive oil sample. The reverse RIP shows a continuous elution of compounds along approximately the first minute of the sample.

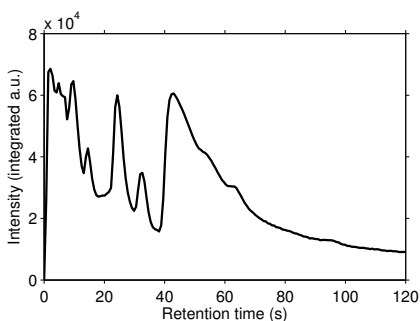


Figure 3.3: Reverse RIP. Analytes eluting from the column will show as a peak in the reverse RIP. The RIP is computed as the integral (from 6.26 to 6.6 ms) of each IMS spectrum and it represents the charge that has not been transferred to other analytes at a given retention time. The RIP's maximum represents the total charge available. By subtracting the charge that has not transferred to the total charge, the reverse RIP is obtained.

The performance of the proposed SW-MCR method has been assessed by comparing the extracted concentration profiles and pure spectra with the ones resolved using conventional MCR-ALS on the whole sample, using the same described pre-processing and imposing the same constraints. Regions with strong co-elution are of particular interest, as for those regions conventional MCR-ALS is not able to resolve all compounds, especially the smallest ones. 22 peaks of the first 100 seconds of the sample were randomly selected (covering higher and lower peak intensities) and we checked the retention time range where each peak had been detected by each method. Table 3.1 shows the actual retention time range of the sample and the one obtained by each method. When the peaks are detected, there is considerable agreement between both methods; however MCR-ALS failed to detect 9/22 of the analyzed peaks.

Figure 3.4 shows a sample region with co-elution and peak intensities of different magnitudes: at retention time 40 s two peaks appear: a peak of 2200 intensity units at drift time 10 ms and a less intense peak of 650 intensity units at 7.8 ms. Close to 50 s a third peak of 230 intensity units appears at 8.7 ms. Given the intensity and large tailing shape of the 10 ms peak, it is reasonable to think that the smaller 8.7 ms peak was co-eluting before its detection, but was being masked by the largest peak due to the charge competition effect. Nevertheless, the difference of the peak intensities detected by the IMS is almost of one order of magnitude.

Table 3.1: Localization of 22 randomly selected peaks from the sample, on MCR-ALS and on SW-MCR deconvolution. The retention time ranges are in agreement on the detected peaks, however MCR-ALS was unable to extract 9/22 peaks.

Peak #	Drift time (ms)	Max. intensity (a.u.)	Retention time range (s)		
			Sample	MCR-ALS	SW-MCR
1	6.45 (RIP)	3951	All	All	All
2	7.30	3936	1-4	0-10	0-4
3	7.60	1002	3-7	2-7	3-9
4	7.75	400	4-10	NF	4-9
5	8.30	782	5-9	4-8	4-10
6	9.10	177	5-12	NF	4-10
7	8.10	695	4-9	0-20	4-10
8	8.78	731	5-12	4-12	5-12
9	8.60	623	4-12	4-7	5-13
10	7.15	474	4-13	3-7	5-15
11	8.90	2100	8-13	8-12	8-13
12	8.10	425	11-19	NF	12-20
13	6.75	490	11-23	NF	12-23
14	10.30	1190	20-27	22-27	21-27
15	7.20	188	21-29	NF	21-28
16	8.20	481	21-32	NF	21-31
17	8.50	390	30-38	30-35	30-37
18	7.80	317	28-40	NF	30-40
19	7.30	445	31-40	NF	32-40
20	9.90	2200	40-55	40-60	40-50
21	8.70	220	50-63	NF	50-56
22	7.65	650	45-80	50-80	55-80



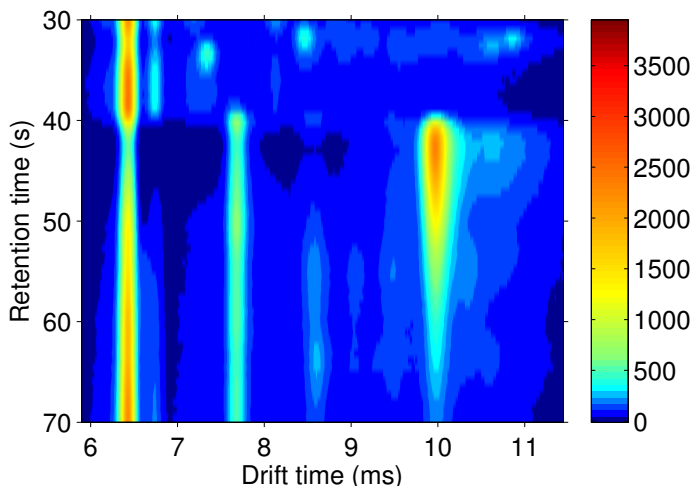


Figure 3.4: MCC-IMS region (contour plot). This region shows co-elution of different compounds: a section of the RIP found at 6.4 ms and three other compounds appear 7.8 ms, 8.6 ms (less intensely) and 10 ms.

Using MCR-ALS in the whole sample, the resolved pure spectra and concentration profiles on the described region are shown in figure 3.5. The only meaningful compound extracted at that retention time region apart from the RIP is the most intense one, found at 10 ms and marked using a wider line. Other compounds appear, some of them can be interpreted as tails or replicas of the 10 ms peak, but they provide no particular meaning so they must be discarded as spurious compounds. Additionally, the concentration profile for the resolved peak shows non-zero concentration in the 20-40 second retention time region, before the compound has eluted.

When using SW-MCR, the pure spectra and concentration profiles for the three peaks on the described region are extracted (see figure 3.6). The computed error bars of the pure spectra and concentration profiles show a high consistency among different window estimations. As expected, the concentration of the largest peak (at 10 ms) is similar to the concentration resolved using MCR-ALS. The peak with lowest intensity (at 8.7 ms) is well resolved too, with a concentration profile one order of magnitude smaller than the largest peak, as expected. The medium intensity peak (at 7.8 ms) is also detected, although its tracking is interrupted in the 47-53 seconds range. This shows a limitation of the proposed technique: Peaks with a constant intensity in the whole window cannot be detected by

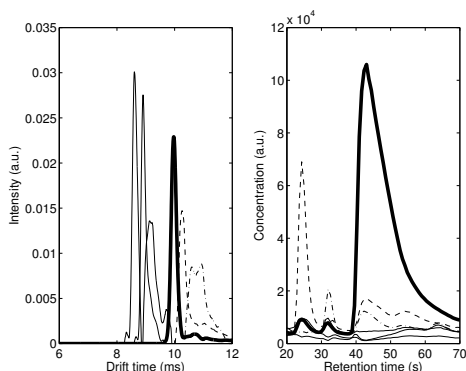


Figure 3.5: Pure spectra and concentration profiles resolved by MCR-ALS. Thick line: main peak resolved. Dashed lines: tails and replicas of the resolved peak. Thin lines: spurious compounds.

SIMPLISMA because the standard deviation of the peak maximum along the window is zero leading to zero purity values. However, the peak is tracked again in further windows once the intensity varies again. As expected, neither MCR-ALS nor SW-MCR were able to deconvolve the 8.7ms peak when it was being completely masked by the 10ms peak.

The SW-MCR technique allows extracting detailed information of the co-elution present in the sample. Figure 3.7 shows the distribution of the resolved compounds along the retention time. Each row represents a tracked compound, showing the retention time region in which it has been detected and deconvolved. For instance, the first row represents the RIP, which is tracked along all the chromatogram. Figures 3.2 and 3.3 showed multiple compounds co-eluting from the column on the first seconds of the analysis. Figure 3.7 confirms that SW-MCR is able to detect them, resolving more than 6 compounds on a single window. As the retention time increases, figure 3.2 shows less peaks co-eluting, and this is reflected on figure 3.7 as the compound overlap decreases. On the first 100 seconds of the sample, SW-MCR was able to track up to 46 compounds revealing the richness of information present in the MCC-IMS samples.

The estimated concentration profile of the RIP is shown on figure 3.8. Retention time regions with lower RIP concentrations indicate regions with intense peaks, or regions with multiple peaks detected, where (almost) all the charge has already been transferred. The recovered concentration profile for the RIP can be compared with the extracted reverse RIP shown in figure 3.3. Figure 3.8 peaks can be matched with figure 3.3 valleys, proving the consistency of our technique.

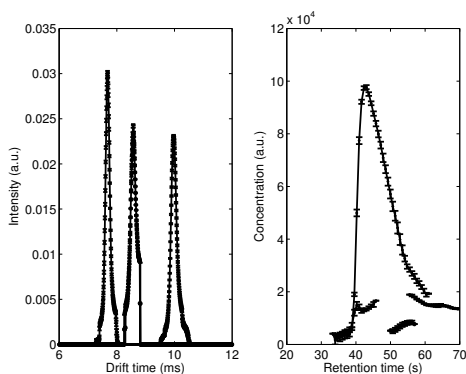


Figure 3.6: Pure spectra and concentration profiles resolved by SW-MCR, deconvolving three co-eluting spectra. The small size of the error bars shows a high consistency among the tracked windows.

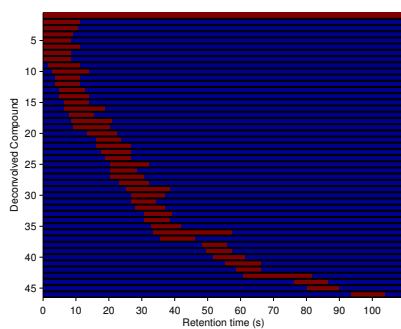


Figure 3.7: Tracked compounds along several windows. Each row represents a different pure spectrum. The first one is the RIP, which is being tracked along all the windows. This plot shows how co-elution of three and more components can be resolved at several retention times.

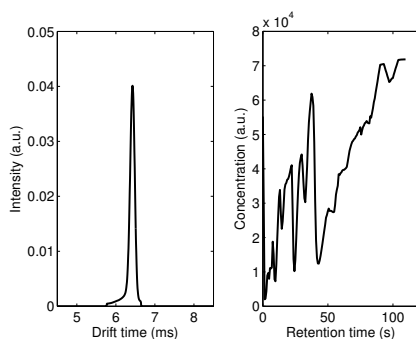


Figure 3.8: The extracted Reactant Ion Peak using SW-MCR. The reverse shape of the concentration profile is comparable to the reverse RIP shown in Fig. 3.3.

The angle threshold used for tracking the compounds rejects spurious peaks that appear on a single window. Figure 3.9 shows the estimated number of components using singular value decomposition (SVD) and the actual number of tracked components for a particular window. Regarding the number of tracked components, all windows are able to track the RIP and therefore the number of tracked components is always greater than or equal to one compound. Most of the windows had none or one spurious compound, although some windows had up to five spurious components which were rejected. This rejection allows us to use lower thresholds on the SVD, as overestimations in the number of components of a window are regulated in the peak tracking step. Lower thresholds on the singular values allow us to detect peaks close to the noise level, as they are consistent across different windows and are not rejected.

Regarding the computational cost of the technique, most of the computing time is spent on the MCR-ALS optimization, as it is an iterative algorithm. For the conventional MCR-ALS, up to 30 iterations are required to reach convergence. For the proposed SW-MCR a maximum of 10 iterations per window were used although for most of the windows 5 iterations were enough for MCR-ALS to converge. In any case, the most expensive part is the (non-negative) least squares estimation required on each concentration profile and pure spectrum estimation. On a 2013 workstation computer with an Intel i7 processor, the conventional MCR-ALS method takes 4.5 minutes to extract the concentration profiles and pure spectra from the sample. This process cannot be parallelized due to its iterative nature. The proposed SW-MCR method requires 1.75 seconds per window. Given that there is a strong window overlap, the overall cost per sample sums up to 22 minutes. Even though the global time is higher, the SW-MCR method is well suited for parallelization, as each window is independent from the others. For

our case, with a window shift of 0.7 seconds, three CPU cores would be needed for a real time application, unfeasible with the conventional algorithm.

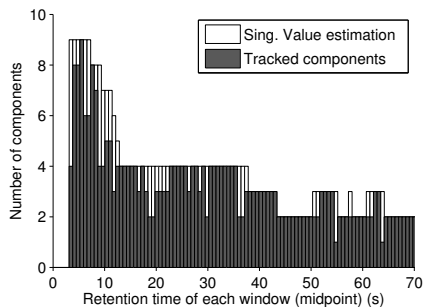


Figure 3.9: Rejection of spurious compounds. In black: the number of tracked compounds for each window. In white: the initial estimation of the number of components using SVD. The difference of both black and white values is the number of rejected spurious compounds.

In summary, a novel technique for improved chemometric resolution of GC-IMS samples has been presented, showing that blind peak deconvolution techniques can be successfully applied to this analytical hyphenated instrument in order to extract features and resolve coelution even when conventional MCR-ALS is unable to discriminate lower intensities from the sample noise.

# Chapter 4

## Applications

This chapter shows two applications of some of the described techniques to real complex samples. The first application shows a classification problem olive oil samples according to their quality measuring their headspace using GC-IMS. The second application focuses on the analysis using GC-MS of the headspace of human urine samples, for trying to discover prostate cancer biomarkers. These two applications from fairly different fields and with different instrumentation highlight the versatility of some of the techniques developed. Even though the instrumentation and the field of application may be different, there is a shared ground for data analysis. While the “one shoe fits all” approach to use the same methodology in all cases is far too generic given the existing differences, exploring data analysis methodologies on different instrumentation can lead to a better understanding of those data analysis methods and provide ideas to better adapt them to specific needs.

### 4.1 Olive oil quality analysis with GC-IMS

The first application is related to fraud prevention in food. Different types of olive oil have different organoleptic characteristics, and therefore different value in the market. Simple and accurate classification of olive oil types is desirable for the industry to reduce costs and prevent fraud. Olive oil can be categorized into “extra virgin” olive oil (EVOO), “virgin” olive oil (VOO) and “lampante” olive oil (LOO), according to their quality.

We are not necessarily interested on a specific peak or analyte that characterizes

each of the classes. Even if that is a possible solution we are looking for a multivariate pattern or fingerprint able to discriminate among the different olive oil categories based on a MCC-IMS analysis of the olive oil headspace.

In contrast to common approaches that reduce the dimensionality of chromatograms into a vector of peak descriptors that serves as input to the multivariate models, in this application we use 2D intensity maps directly for modelling. With this methodology we avoid peak picking methods that may yield errors due to previous peak area definitions, that happen more often when there are high degrees of coelution. Peak mismatching errors and the need to impute missing values in peak tables are also issues avoided, as full-spectra alignment techniques are used instead.

The complete data matrix is carried until the model is built, reducing the complexity of data pre-processing and therefore minimising the possibility of introducing artifacts in the analysis. The workflow for preprocessing and processing MCC-IMS data along with an example of data acquired to differentiate between olive oil categories is given in this section.

### 4.1.1 Experimental Protocol

We used a set of olive oil samples of three different categories. Each sample was analyzed twice, resulting in a set of 216 MCC-IMS chromatograms that corresponded to 3 different types of olive oil: 92 chromatograms were extracted from “extra virgin” olive oil (EVOO) samples, 64 from “virgin” olive oil (VOO), and 60 from “lampante virgin” olive oil (LVOO).

Software LAV version 1.5.2 beta from G.A.S was used for data acquisition, collecting spectra in positive ion mode. More information on the sample protocol can be found in ref (Garrido-Delgado et al., 2012).

The samples were stored at 4 °C in opaque glass containers and then analyzed by means of a MCC-IMS instrument (FlavourSpec®) from Gesellschaft für analytische Sensorysysteme mbH (G.A.S.), Dortmund (Germany) according to (Commission and others, 1991). The instrument was equipped with an autosampler device (CTC-PAL, CTC Analytics AG, Zwingen, Switzerland). An aliquot of 1 g of olive oil sample was placed in a 20 ml glass vial and it was hermetically closed with a magnetic cap. Then, each sample was heated at 60 °C during 10 min in order to generate a headspace into the vial. After this, 100  $\mu$ L of gaseous sample from headspace were injected by the autosampler device into the heated splitless injector (80 °C) of the MCC-IMS instrument.

After injection, gaseous sample was dragged to a non-polar OV-5 MCC (20

cm long, consisting of approx. 1,000 parallel glass capillaries) which was thermostated at 30 °C. Into the MCC, analytes of the sample were separated and subsequently carried to the ion mobility spectrometer for its detection and quantification. In the spectrometer, analytes were delivered into the ionization chamber for ionization by a Tritium source (6.5 keV, 300 MBq). Subsequently the ions were introduced to the drift tube (6 cm of length) with a grid pulse width of 100  $\mu$ s where they traveled under a constant electrical field (350 V/cm) and constant temperature (60 °C) to reach the detector. For noise removal purposes, the instrument was set to record a spectrum every 32 scans, that were averaged. One scan was acquired every 21 ms, leading to an effective sampling period in the retention time of 672 ms. The sampling frequency in the drift time axis was of 150 kHz.

Given the short amount of time the ion shutter is open with respect to the time a spectrum acquisition lasts, only a small percentage of the ions produced are actually measured (Karpas, 2000), and there is a reservoir of ions remaining in the ionization region, that appear in subsequent scans. This causes a characteristic peak broadening effect in retention time that does not appear on other conventional chromatographies. While reducing the number of scans used for averaging would increase the retention time sampling frequency, the obtained peaks would not be thinner but the signal to noise ratio would be smaller, due to the lack of averaging.

### 4.1.2 Data analysis methodology



Figure 4.1: Data analysis flow chart for olive oil discrimination

Figure 4.1 shows the steps of the MCC-IMS data processing workflow presented in this chapter. The procedure begins with the described acquisition of MCC-IMS data from different olive oil samples and corresponding metadata (time of analysis and olive oil category). Next, samples are preprocessed to reduce noise, correct baseline, and align spectra. The preprocessed spectra are unfolded and passed to a Partial Least Squares - Discriminant Analysis model, with the corresponding labels for each sample (EVOO, VOO or LVOO). The number of latent variables of the PLS-DA model is chosen using internal validation, and the performance is evaluated with external validation.



**For preprocessing**, the Savitzky Golay filter and the psalsa baseline estimation methods described in chapter 2 are used. For the spectral alignment, the multiplicative correction of the drift time was aligning well the RIP peaks, but other peaks in the spectra needed further correction. To align completely the spectra, the alignment was refined using a warping alignment method based on the compression and expansion of peaks named Correlation Optimized Warping (COW) (Tomasi et al., 2004). It is based on a lineal warping of the drift time of spectra segments, to maximize the correlation between the reference and the sample to align. As this dataset did not have any internal standard nor calibrant, a simple linear model was used to align the retention times.

**Regarding the classification**, a Partial Least Squares Discriminant Analysis (PLS-DA) model was used. PLS-DA is a widespread supervised classification technique that aims to reduce the dimensionality of the input data (the MCC-IMS spectra) into a linear subspace of much smaller dimension (to be optimized). This subspace is chosen such as it maximizes the variance of the spectra and its covariance with the class labels. A linear regression of the input subspace versus the class labels is estimated, and prediction of classes on new samples can easily be computed using a linear projection.

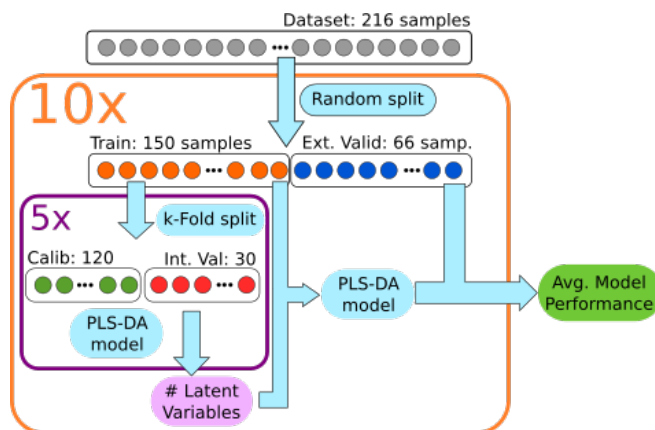


Figure 4.2: We performed double cross-validation to evaluate the classification accuracy of the model. The complete dataset was randomly divided in training (150) and test samples (66). Internal validation was carried out by means of 5-fold cross-validation and external validation was evaluated with test samples. The process was repeated 10 times to provide error of the accuracy

**To validate the model** and to estimate its performance, double cross-validation was used. The methodology is represented in Figure 4.2. The average model

performance is estimated as the mean of the performances of several (10) models, where each of them has been trained with a random split of the data, using 150 samples for training and 66 samples for test. The samples in the training subset are split into five folds and a k-fold approach is used to estimate the optimal dimensionality of the PLS subspace, using four folds for model calibration and the remaining fold for internal testing, choosing the number of latent variables. The dimensionality of the linear subspace is known as the “number of latent variables” of the model, as each dimension of this subspace combines common variability correlated with the labels. For each of the k-folds, we explored the range from one to twenty latent variables, using the classification rate of the internal-validation samples as a figure of merit for model performance. A higher number of latent variables also means higher model complexity that can hinder model interpretation and be more prone to overfitting. A model with too few latent variables may not be able to capture the different sources of correlation (typically several analytes) offering bad prediction capabilities.

### 4.1.3 Results

#### 4.1.3.1 Preprocessing

Figure 4.3 shows as an example an IMS spectrum extracted at retention time 2 min 20 s. The Savitzky-Golay filter used for denoising has a window length of 0.1 ms. Compared to the full width at half maximum of 0.15 ms of the IMS peaks, the chosen window length is large enough to reduce noise but still small enough to prevent the distortion of peak shapes. After removing high frequency noise, we estimate and compensate the baseline: The *psalsa* algorithm is set to  $\lambda = 100\,000$ ,  $p = 0.01$  and the default  $k = 5\% \max(I)$  of the maximum intensity in the spectra. *psalsa* converges rapidly, allowing to estimate the baseline just after five iterations. A manual supervision of few representative spectra is useful to check that the election of parameters is robust throughout the dataset.

The effect of the drift time alignment is represented in figure 4.4, the reactant ion peaks are located at the same drift time while other peaks (e.g. at 7.4 ms or at 8.1 ms) are also properly aligned.

On the other axis, the result of the linear retention time alignment is represented in figure 4.5. While more flexible retention time corrections have been presented in previous chapters, these corrections required the use of calibrants that were not available in this study.

A flexible alignment technique is able to cope with all the non-linear changes and provide a better fit to each of the peaks if they are well detected. However, it

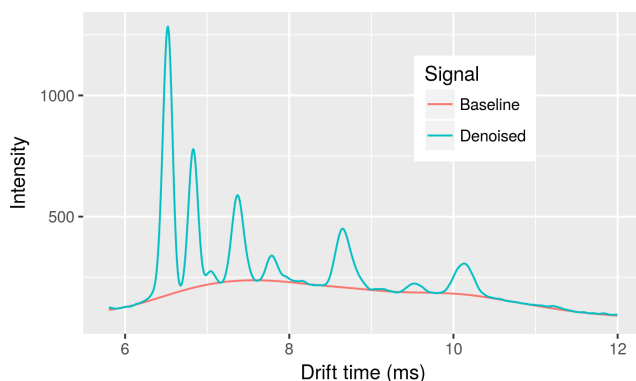


Figure 4.3: Denoised IMS spectrum using a Savitzky-Golay filter and its estimated baseline (using the *psalsa* algorithm)

is also easier to mismatch peaks with flexible techniques. A reasonable midpoint consists of starting with more rigid alignment techniques, finding a coarse solution and then apply a more flexible technique, with constraints that minimize possibly unwanted distortions.

#### 4.1.3.2 Model training and accuracy

The preprocessed spectra were used to train and test the PLS-DA models as described in 4.1.2. Figure 4.6 shows the classification accuracy in internal validation data as a function of the number of latent variables. The accuracy of the model increases with the number of latent variables until it reaches a plateau. If the number of latent variables kept increasing, the model would start to overfit the training data and this would have resulted in accuracy decay when evaluated with the internal validation data. To provide further details of the impact of the preprocessing in this dataset, we explored how each of the preprocessing steps affected the classification accuracy of the models. The whole analysis was repeated, removing one of the preprocessing steps on every repetition. Figure 4.6 also shows how, for this particular dataset the baseline correction had significant impact on the final accuracy, the alignment correction had a minor but still relevant impact and the noise reduction did not contribute to any improvement for this dataset.

The models, built with 6 latent variables, were finally evaluated with the external validation samples, resulting in an accuracy of  $85\% \pm 5\%$ .

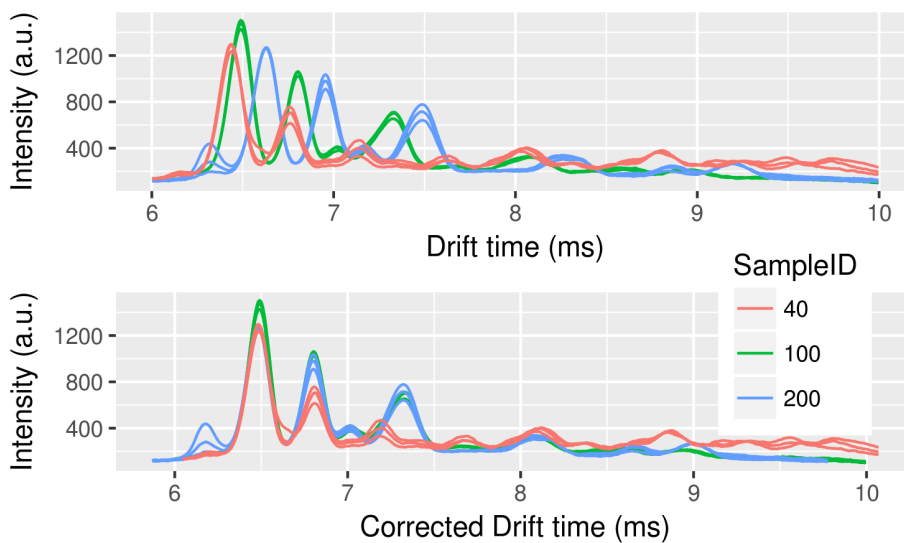


Figure 4.4: RIP alignment multiplicative correction for three spectra measured at retention time 1 min, on three samples one of each class. The correction factor was less than 2% in all cases.

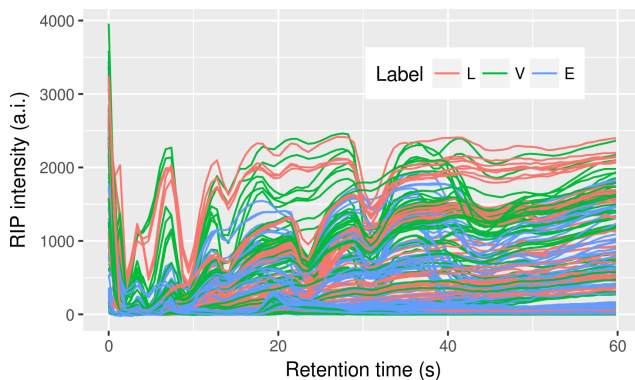


Figure 4.5: Retention time alignment for olive oil samples

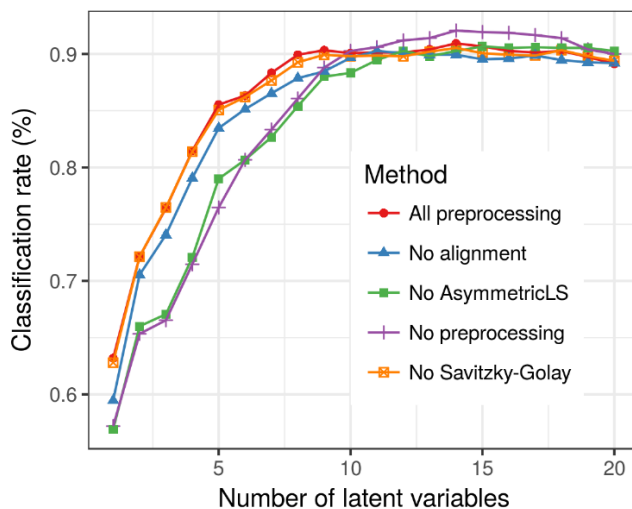


Figure 4.6: Classification rate vs model complexity. The number of latent variables was selected evaluating the classification accuracy in internal validation

An in-depth look at the PLS-DA models provides information on the relevant areas of the MCC-IMS samples for their correct classification. Figure 4.7 shows data samples reduced to only two latent variables. Training samples were used to obtain the directions of the first two latent variables. When test samples are superimposed on the same plot one can conclude that the model exhibits good generalization since both sets (training and test) cover the same regions of the new space for the different type of olive oils. Additionally, one can observe that LOO samples appear further apart than EVOO and VOO, indicating that the identification of LOO is easier than the classification of the other two types of samples. This is particularly important since, unlike the other two types, LOO is not certified for human consumption and has lower market value. Finally, EVOO samples tend to exhibit higher scores on the first latent variable. When exploring the loadings for the first latent variable (see Figure 4.8) the relevant areas to identify EVOO can be identified. Samples with higher intensities than the mean in the purple regions or lower than the mean in the red regions will likely correspond to EVOO.

This exemplification of a methodology for the analysis of MCC-IMS samples using multivariate methods does not depend on peak integration techniques. The methodology has a strict validation scheme to prevent overfitting the model to calibration data, to optimize the model metaparameters and to estimate the

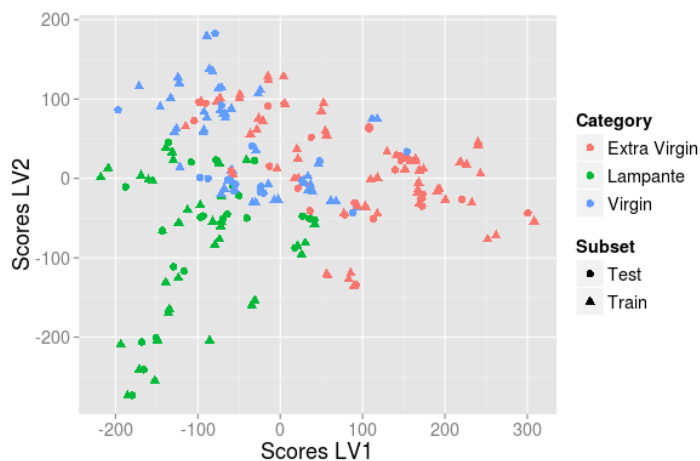


Figure 4.7: Scores for the first two latent variables of the training set. The same projection is used for the test samples. EVOO samples tend to exhibit higher scores on LV1.

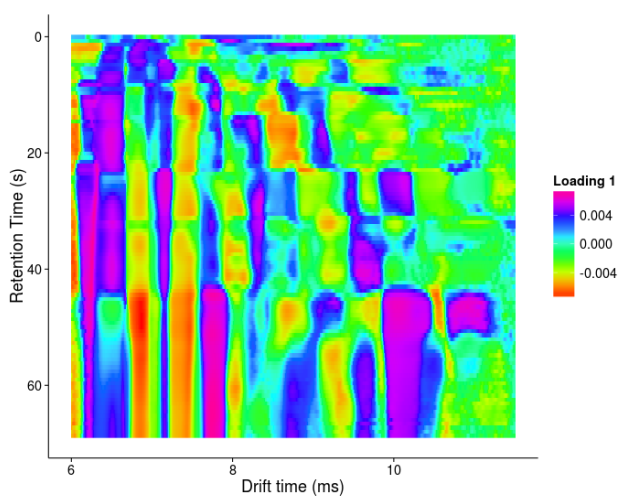


Figure 4.8: Loadings for the first latent variable. Samples with higher (lower) values than the mean in purple (red) areas will appear with high values for the first latent variable.

performance of the classification.

By applying this methodology to a specific olive oil analysis dataset, we have been able to explore the impact of each of the preprocessing steps on the final classification. For this specific problem, we have shown that proper baseline estimation and alignment are relevant for the improvement of the predictive accuracy of the model, however denoising was not a key issue in this scenario.

## 4.2 Prostate cancer biomarker discovery

The second highlighted application is focused on the analysis of the headspace of human urine samples, using Gas Chromatography - Mass Spectrometry, with the goal of detecting biomarkers able to discriminate subjects suffering of a prostate cancer.

Prostate cancer diagnosis is nowadays confirmed through a prostate biopsy. Biopsies are invasive and uncomfortable for the patient, as well as expensive. Biopsies are performed typically after a positive result of the Prostate Specific Antigen (PSA) blood test that acts as a screening method for the prostate cancer diagnosis. However PSA is a controversial test due to the high number of false positives results, mainly related to non-cancerous conditions such as the Benign Prostatic Hyperplasia (Thompson et al., 2004). Having only a positive result in 30% of the biopsies indicates that the other 70% were not necessary, and therefore that finding specific non-invasive biomarkers for prostate cancer would reduce the number of biopsies performed, saving patient discomfort, time and resources.

The analysis of Volatile Organic Compounds in urine had in (Mills and Walker, 2001) one of its major contributions, listing 103 compounds. More recent reviews (de Lacy Costello et al., 2014), report up to 279 VOCs for urine, stating that this number is not larger due to the low concentrations of VOCs in urine (see table 1.1 from chapter 1). In (Cornu et al., 2011), the authors reported progress in the detection of prostate cancer using dogs sniffing urine. A bit later, in (Khalid et al., 2015), the authors have reported the detection of potential biomarkers, with a classification accuracy between 60% and 70%.

### 4.2.1 Methods

The urine samples were provided by the Vall d'Hebron hospital and stored at -80 °C. Each sample was analyzed using Gas Chromatography - Mass Spectrometry. The chromatograph was a Trace GC Ultra from Thermo Fisher, equipped

with an autosampler device. An aliquot of 3 ml of urine, either in a neutral or a basic pH medium was placed in a 20 ml glass vial and it was hermetically closed. The injector was set to 220 °C. The analysis lasted 27 min, with a temperature ramp from to °C. The GC was connected to a single quadrupole MS through a transfer line at 260 °C. The MS was set to scan the mass range from 28 to 350 m/z.

Samples were randomized before analysis to prevent a date/condition confounding factor. The dataset included a blank sample every 4 analysis, to assess that there was no cross-sample contamination. External calibrants (diethyleter, etyhl decanoate and bromoform) were also sampled after each blank sample, and used as references in retention time alignment to correct any possible experimental drift. These particular calibrants were chosen because they presented different retention times at 5, 14 and 24 minutes allowing us to detect drifts on the beginning, the centre and the end of the chromatograms respectively.

### 4.2.2 Data analysis

The preprocessing of the urine samples consisted of the following steps:

- Blank subtraction
- Baseline removal, using the *psalsa* algorithm.
- Denoising, using a *Savitzky-Golay* filtering.
- Peak extraction and Peak integration, using the *Bieller-Biemann* deconvolution.
- Alignment, using the *Robinson-Souza* algorithm (Robinson et al., 2007).

These steps were integrated with the PyMS toolbox (O’Callaghan et al., 2012), customizing the denoising and baseline correction steps to integrate the Savitzky-Golay filtering and the *psalsa* baseline estimation method.

Right after the sample was binarised in a matrix, its associated blank was subtracted from it, in order to remove the effects of the compounds present in blank samples, coming from column or contamination. After blank subtraction, the baseline of the samples was subtracted using *psalsa*, with  $\lambda = 100000$  and  $p = 0.01$ . The effect of the baseline removal on the whole sample is shown on figure 4.9. After the baseline removal a Savitzky-Golay filter of 7 points long and second order was used to remove noise.

A peak is represented by a retention time, an intensity and a mass spectrum. From each sample, a list of peaks was compiled using the Biller—Biemann algorithm using 5 points and 10 scans for the deconvolution. Peak areas were integrated using only the top most common ions, as marginal ions account for



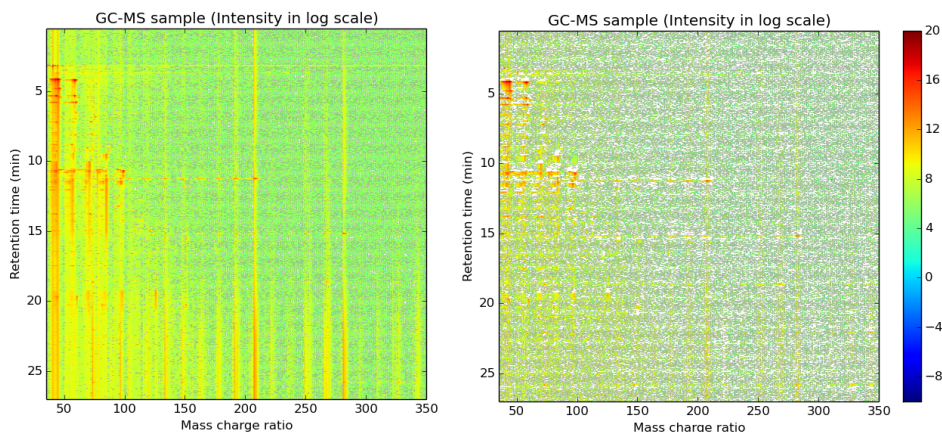


Figure 4.9: Baseline subtraction effect. On the left: sample before the subtraction. On the right: the same sample after baseline subtraction.

sample noise. To avoid false positives in peak detection, a threshold was imposed to ignore small peaks corresponding to noise. The results of the peak detection are represented over the Total Ion Chromatogram on figure 4.10.

The final step to obtain a matrix of features is to compare the extracted peaks among the different samples, and cluster the peaks according to their similarity so peaks from different samples corresponding to the same analyte are clustered. The algorithm to perform the peak alignment is the Robinson-Souza method described at (Robinson et al., 2007). In summary, a similarity score between two peaks  $i$  and  $j$  is defined as:

$$P(i, j) = S(i, j) \exp\left(-\frac{(t_i - t_j)^2}{2D^2}\right) \quad (4.1)$$

where  $S(i, j)$  is the dot product of the respective mass spectra and  $t_i$  and  $t_j$  are the retention times of the respective peaks.  $D$  is a retention time tolerance parameter that was set to 5 s.

Peak lists from different samples are pairwise compared using the similarity score and are aligned using dynamic programming with the Needleman-Wunch algorithm, similarly as done in sequence alignment. A gap penalty of 0.35 was used.

All the peaks from different samples that are clustered together belong to the same compound. A mass spectrum representative of that compound is computed

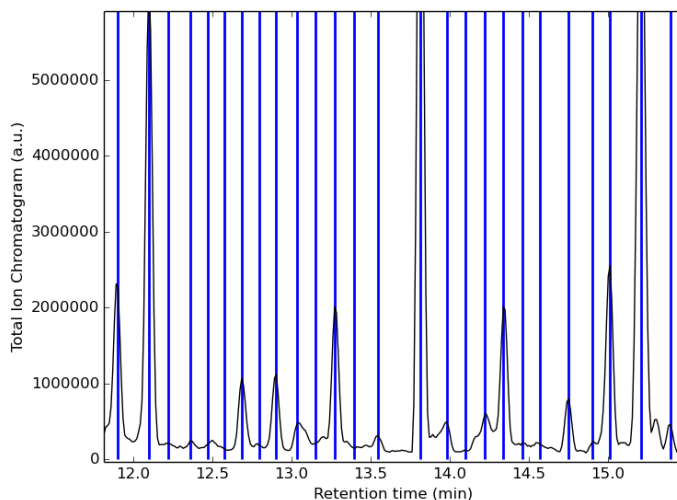


Figure 4.10: Peak detection in GC-MS samples. Fragment of a retention time with the detected peaks highlighted.

as the average mass spectrum of the peaks in the cluster. Figure 4.11 gives an example of a comparison between the representative mass spectrum of each of the compounds with the mass spectrum of methanethiol, extracted from the NIST reference library, showing how methanethiol has been successfully extracted and clustered from the samples. The alignment procedure creates a final matrix for further data analysis with as many rows as samples are analysed and as many columns as peak clusters have been formed. If a particular sample does not have a compound, then a NaN (Not a Number) value is imputed in the matrix.

After discarding compounds present in less than 60% of the samples, 74 and 71 compounds were found at urine samples in a basic pH and in a neutral pH respectively.

The matrix of peaks and samples can be used for further data analysis. To start, we opted for applying a Wilcoxon test (Bauer, 1972) and a Rank Products test (Breitling et al., 2004). This univariate analysis did not reveal any compound as significant, indicating that there is not an individual biomarker able to discriminate between cancer and control patients. Nevertheless, figure 4.12 shows a boxplot for the peaks with larger differences between both groups. This lack of finding was discouraging, and further multivariate analysis were tried without

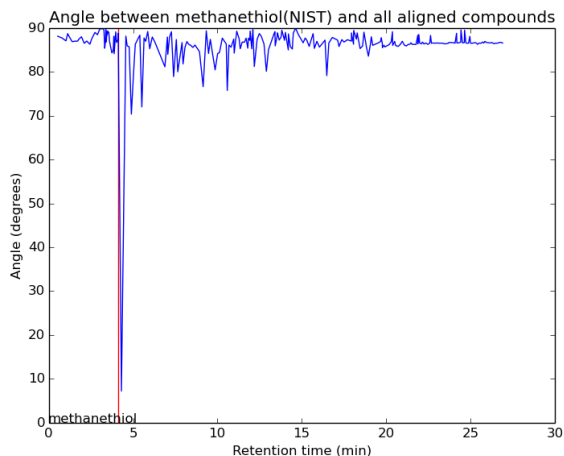


Figure 4.11: Angle between methanethiol mass spectrum extracted from the NIST and each of the compounds automatically extracted from the samples. Angle is above 70 degrees for all the compounds except for the actual methanethiol, where the angle decreases below  $10^\circ$ .

luck.

To further assess that the problem was not found in our data analysis methodology, and for the sake of curiosity, a master student was put to work on data analysis for GC-MS using this same dataset. His approach, based on other tools such as XCMS (Smith et al., 2006) and MzMine (Pluskal et al., 2010), obtaining similar results reported at (Macías, 2017).

We therefore assume that the problem must be either in our instrumentation, in our methodology, or in the reproducibility of the analysis. Besides this, this application shows some of the tools we described in the thesis and it is a sample of how hard getting reproducible results can be in this field.

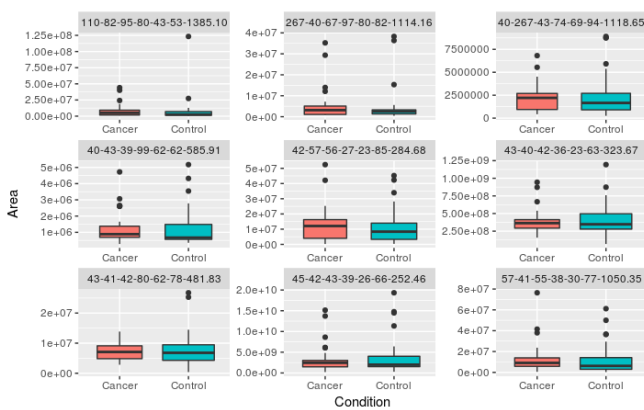


Figure 4.12: Boxplot of the peaks with largest differences in the medians for both cancer and control patients



# Chapter 5

## Conclusions

This thesis studies data analysis tools and algorithms for hyphenated analytical chemistry instrumentation, focusing on Gas Chromatography - Ion Mobility Spectrometry.

- The study in section 2.2 of several baseline estimation methods for GC-MS and GC-IMS samples has exposed a limitation of the Asymmetric Least Squares technique for the baseline correction of samples with large dynamic range. The modification of the weights in the ALS algorithm so they depend on the magnitude of the residual has been effective to overcome the described limitation, improving the baseline estimation both in synthetic and real samples.
- The characterization of the misalignments from section 2.3 performed on a GC-IMS dataset that spans more than 10 months has shown that first order drift time misalignments can be corrected with a linear warping, having a slope correction factor smaller than 3%. The correction factor, as seen on figure 2.14 is clustered by the day of the analysis indicating a correlation of intra-day misalignments in drift time. However, retention time misalignments benefit more from non-linear corrections, so we have explored how monotonic cubic splines improve linear warping methods thanks to their flexibility. Based on those results, we have suggested an optimal time span between two consecutive external calibrant measurements of two days. In spite of the proposed correction based on external calibrants, the use of internal calibrants is still convenient for the correction of retention times, in order to reliably align each sample, covering for not systematic intra-sample variations.

- A novel technique for improved chemometric resolution of gas chromatography ion mobility spectrometry samples has been presented, showing for the first time that blind peak deconvolution techniques can be successfully applied to MCC- IMS instrumentation.
- The SW-MCR technique has been tested on olive oil headspace samples. The samples analyzed present a strong co-elution of the individual chemical components, as shown by the wide peaks in the reverse RIP ion chromatogram. Coeluting peaks are not properly resolved using conventional MCR-ALS methods, as peaks of lower intensities cannot be discriminated from the sample noise. Additionally, spurious compounds appear requiring supervision of the results.
- Using the proposed SW-MCR method, we were able to deconvolve the pure spectra and concentration profiles of most of the peaks, even the ones with lower intensities, rejecting spurious solutions automatically.
- Further work in the SW-MCR area can be oriented to improve the initial estimation of the concentration profiles and pure spectra, in order to overcome the limitations in SIMPLISMA to resolve peaks of constant intensity along the entire window.
- The computational cost of the SW-MCR technique is higher than the cost of the conventional MCR-ALS method, mainly because of the higher window overlap. However, our method can be easily parallelized, making our method more scalable and even suitable for real time applications: with our SW-MCR the retention time windows can be analyzed as they are acquired from the instrument, while the whole sample matrix is needed if a full matrix deconvolution with MCR-ALS is done.
- We have shown the application of some of these signal processing tools to the classification of olive oil samples according to their quality, using some of the described preprocessing methods to extract the chemical fingerprints and train a classifier, using state of the art validation techniques.
- Some of the developed methods were also applied to GC-MS data, looking for prostate cancer biomarkers in urine volatiles. Even though we identified analytes with potential of being biomarkers according to the literature, we did not obtain good prediction capabilities.

We hope that the implementation of these techniques in an open source repository will facilitate its adoption and benchmarking by others.

# Appendix A

## Resum de la tesi

### A.1 Introducció

L'objectiu d'aquesta tesi és el desenvolupament d'algoritmes per l'anàlisi de mostres complexes en fase gasosa fent ús d'instrumentació acoblada, en particular de Cromatografia de Gasos - Espectrometria de Mobilitat d'ions (GC-IMS) i la seva aplicació en mostres complexes reals.

Explorarem les tècniques existents per l'anàlisi de dades tant per GC-IMS com per altres tècniques similars o properes, i proposarem millores modificacions i adaptacions d'algoritmes existents per tal que s'ajustin millor a les característiques del GC-IMS. Avui dia, bona part de les anàlisi estadístiques de dades de GC-IMS es fonamenten en l'ús de sistemes propietaris tancats, proporcionats pel fabricant de l'instrument o per una tercera part (p.ex. VisualNow<sup>1</sup>), o fan ús de tècniques escampades a múltiples paquets de programari i articles. Els algoritmes desenvolupats en aquesta tesi seran publicats en una caixa d'eines oberta i modular, que podrà estendre's i ser reutilitzada per la comunitat.

L'espectrometria de mobilitat d'ions és una eina analítica utilitzada cada cop en més camps: En l'àmbit de la seguretat, l'IMS s'utilitza a diari en aeroports per la detecció de drogues i explosius (Eiceman et al., 2013). Si bé inicialment els usos de l'IMS estaven centrats principalment en la detecció d'agents de guerra química, explosius... avui dia han proliferat d'altres aplicacions (Armenta et al., 2011) com ambientals (Karpas et al., 1991), (Baumbach et al., 1993); industrials (Budde, 1995); estudis biomèdics (Westhoff et al., 2010), (Baumbach, 2009); o

---

<sup>1</sup><http://www.bs-analytik.de/en/products/software-vocan-visualnow.htm>



qualitat alimentària (Vautz et al., 2006c), (Garrido-Delgado et al., 2012) entre d'altres. El seu ús com a eina de recerca per l'anàlisi de mostres biològiques o de ciències de la vida, en particular per la detecció de compostos orgànics volàtils (VOC).

L'anàlisi de mostres complexes requereix de fer ús de progressos recents en instrumentació acoblada (Sarker and Nahar, 2012). Si bé l'IMS no seria una eina adequada per l'anàlisi d'aquest tipus de mostres a causa de la seva poca selectivitat, una pre-separació dels analits de la mostra fent ús de cromatografia de gasos permet superar aquesta limitació, a costa de sacrificar portabilitat i velocitat d'anàlisi. A més, la pre-separació dels analits aporta informació que pot ser útil per la identificació dels analits de la mostra. Al generar una quantitat de dades més grans i més complexa amb aquest instrument acoblat cal fer ús d'algoritmes específics capaços d'extreure tota la informació de la mostra.

Per tal de proporcionar una bona base per a la descripció dels algoritmes proposats, aquesta introducció descriurà breument mostres típiques complexes que són rellevants en aplicacions de GC-IMS. Seguirem explorant la instrumentació analítica, enfocant-nos en els instruments d'interès (IMS, GC-IMS).

Pel que fa les mostres típiques dels àmbits de ciències de la vida, en el camp biomèdic es treballa amb el que s'anomena l'anàlisi del "volatoloma humà". El volatoloma consisteix en el conjunt de compostos volàtils que emanen dels diferents fluids del cos humà. (de Lacy Costello et al., 2014) donen un compendi de la quantitat de compostos orgànics volàtils que s'han arribat a comptar en la literatura, mostrant la complexitat inherent a les mostres, resumit a la taula A.1.

Table A.1: Nombre de VOCs per fluid corporal, segons (de Lacy Costello et al., 2014). Cal deixar clar que no és el nombre total de VOCs existents, ja que per exemple a la orina se sap que n'hi ha més per identificar, presents en molt baixes concentracions

Fluid corporal	Nombre de VOCs
Alè	872
Saliva	359
Sang	154
Llet	256
Secrecions de la pell	532
Orina	279
Fems	381

L'anàlisi del volatoloma no està complet, i hi ha problemes oberts que cal resoldre per transferir els resultats de la recerca a la clínica, entre ells un problema de poca reproducibilitat en estudis pre-mèdics (Begley and Ioannidis, 2015). Disposar d'eines d'anàlisi de dades que segueixin les millors pràctiques (Broadhurst and Kell, 2007) és un pas per tal de millorar aquesta reproducibilitat, i aquesta tesi segueix aquest camí. Així, en aquesta tesi veurem l'aplicació d'alguns dels algoritmes desenvolupats a un conjunt de dades de volàtils d'orina mesurats amb GC-MS, amb l'ànim de descobrir biomarcadors de càncer de pròstata.

Un altre camp que es beneficia de l'anàlisi de compostos volàtils és el de la indústria alimentària, on diferents anàlisi són habituals pel control de qualitat d'aliments i per detectar aliments en un estat de conservació deteriorat. Si bé les eines analítiques habituals requereixen preparació de la mostra, temps i són cares (Vautz et al., 2006c), en la darrera dècada han aparegut mètodes alternatius que complementen aquestes tècniques o les poden reemplaçar. Per exemple, l'ús de IMS per detectar si el menjar es fa malbé (Raatikainen et al., 2005), el control de qualitat de vins (Karpas et al., 2012b) i la prevenció de frau en oli d'oliva (Garrido-Delgado et al., 2012) en són algunes aplicacions.

Aquesta tesi mostra l'aplicació d'alguns dels algoritmes desenvolupats a un conjunt de dades de GC-IMS de l'anàlisi d'olis d'oliva, per la seva classificació d'acord amb la seva qualitat

### A.1.1 Instrumentació

L'espectrometria de mobilitat d'ions és una tècnica analítica per caracteritzar substàncies químiques basant-se en la velocitat d'ions en fase gasosa en un camp elèctric (Eiceman et al., 2013). És capaç de detectar traces d'alguns compostos volàtils, arribant a concentracions del rang de ppb. L'anàlisi és ràpid (un espectre s'adquireix en desenes de milisegons) i l'instrument té una selectivitat moderada.

L'IMS es divideix en una regió d'ionització, una graella d'obtenció i la regió de deriva. Tal i com es mostra a la figura A.1, la mostra entra a la regió d'ionització (en aquest cas després d'haver passat per un cromatògraf). Les molècules de la mostra s'ionitzen i aquests ions passen a la regió de deriva quan la graella d'obtenció s'obre, viatjant a través del tub de deriva, que consisteix habitualment d'una sèrie d'anells metàl·lics que creen un camp elèctric constant. Els ions viatgen pel tub de deriva impulsats per una força que depèn de la seva massa i càrrega, col·lisinant amb les molècules neutres del gas de deriva (aire o  $N_2$  habitualment) que flueix en la direcció oposada. Depenent de la secció eficaç entre els ions i el gas, hi haurà més o menys col·lisions afectant al temps que trigaranà cada molècula a recórrer el tub de deriva. Així els ions se separen d'acord

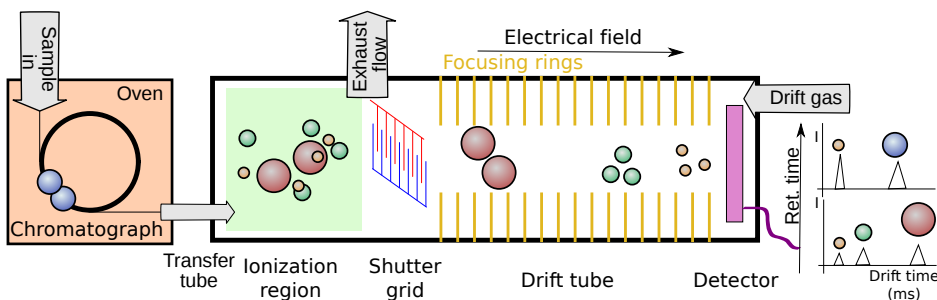


Figure A.1: Diagrama d'un cromatògraf de gasos acoblat amb un IMS

a la seva mobilitat elèctrica, que depèn de la seva massa, forma i càrrega d'entre altres factors, arribant a un detector al final del tub de deriva. El temps que triga un ió a travessar el tub es coneix com a temps de deriva. La relació per a baixos camps elèctrics entre la velocitat mitjana de l'ió i el camp elèctric aplicat és lineal  $v_d = KE$ , on  $K$  és la mobilitat elèctrica. D'aquesta mobilitat elèctrica pot derivar-se la mobilitat elèctrica reduïda, que en compensa les variacions causades per la temperatura i la pressió.

L'ús de determinades fonts d'ionització genera uns ions reactants (habitualment ions hidroni  $H_3O^+$  o ions amoni  $NH_4^+$ ) que són els que interactuen amb la mostra. En cas d'absència de mostra, els ions reactants viatgen sols pel tub de deriva donant lloc al pic (o pics) dels ions reactants (RIP). Aquest pic és especialment útil per les anàlisi en tant que com veurem pot fer-se servir de referència per alinear els diferents espectres. La quantitat d'ions reactants disponible és finita, cosa que fa que en barreges complexes només els compostos que tinguin més afinitat amb els ions reactants hi reaccionin, donant lloc a un efecte de competició per la càrrega d'aquests ions i per tant a no linealitats.

Per tal de millorar la selectivitat de l'IMS, i també per reduir el nombre de compostos presents al mateix temps a l'àrea d'ionització, és possible acoblar-ne a l'entrada un cromatògraf de gasos, que permet separar els compostos d'una mostra en funció de la seva afinitat i capacitat de ser retinguts per una columna cromatogràfica. Per la capacitat de fer cromatografies ràpides i per poder proporcionar fluxos elevats, habitualment s'utilitzen unes columnes multicapil·lars (MCC) que a diferència de les columnes habituals estan formades per un paquet de capil·lars en paral·lel. En una cromatografia el temps que triga un compost a travessar la columna i eluir-ne s'anomena temps de retenció.

La figura A.2 mostra una mostra de MCC-IMS d'oli d'oliva. Els pics d'MCC-IMS són més amples que els típics pics de cromatografia i se'n poden veure uns

quants eluint al mateix temps de retenció, fenomen conegut com co-elució. El RIP descrit anteriorment es pot veure al llarg dels temps de retenció a un temps de deriva proper als 6ms. Veiem també com al eluir altres compostos la intensitat del RIP decreix.

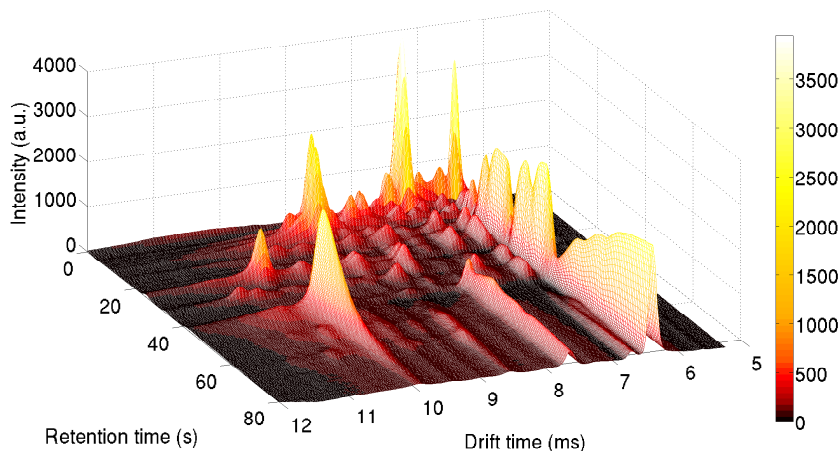


Figure A.2: Regió d'una mostra MCC-IMS d'oli d'oliva. El pic d'ions reactants (RIP) s'observa als 6ms així com es pot veure també el fenomen de la coelució.

Havent vist el tipus de mostres amb les quals treballarem i la instrumentació existent, tenim una base per la discussió dels algorismes existents per l'anàlisi de dades més adient.

## A.2 Preprocessat

El preprocessat és la millora de les dades en cru donades per l'instrument fent ús de filtres i de tècniques de processat de senyal per eliminar soroll i artefactes de les dades, corregint derives instrumentals i variacions de la línia de base. L'objectiu és anar de les "dades en cru" a "dades netes", llestes per fer anàlisi de dades i modelat. L'increment de la quantitat i complexitat de les dades en cru fa més necessari dedicar-hi més temps i esforços a netejar-la, per assegurar-nos que la informació extreta posteriorment té sentit i és d'alta qualitat. Si el preprocessat no es fa amb cura poden aparèixer artefactes i variacions no desitjades (en comptes de ser eliminades!) i això pot afectar a les anàlisi posteriors (Engel et al., 2013).

Per les dades de GC-IMS, els principals problemes que han de ser tractats són el soroll de les mesures, la línia de base, els desalineaments i la normalització. De tots aquests, en aquest resum ens centrarem en la línia de base i l'alineament, que són els dos camps on s'han fet les contribucions més rellevants.

### A.2.1 Correcció de la línia de base

Les tècniques d'estimació de la línia de base s'utilitzen habitualment per corregir els efectes a llarg termini de contaminació o de degradació instrumental i són essencials per poder fer una integració acurada de l'àrea dels pics de la mostra.

Existeixen moltes tècniques per estimar la línia de base i corregir-la, com per exemple mètodes basats en ajustos polinomials (Salit and Turk, 1998), en mínims quadrats pesats (Eilers, 2003), (Zhang et al., 2010), (Peng et al., 2010) o mètodes basats en wavelets (Shao et al., 2003). Alguns dels mètodes requereixen que l'usuari defineixi regions sense pics per estimar la línia de base. Altres mètodes aproximen la línia de base de forma iterativa, mirant de detectar i rebutjar les regions amb pics que no pertanyen a la línia de base

En aquest treball es proposa una modificació del mètode de mínims quadrats asimètrics (ALS) proposat a (Eilers Paul H. C., 2005). Hem trobat que ALS té un comportament esbiaixat en la presència de pics intensos (en relació al nivell de soroll). Aquests pics intensos es troben habitualment a mostres de GC-MS i també GC-IMS.

A (Newey and Powell, 1987). ALS va ser presentat per tal de construir tests estadístics en l'àmbit de la econometria. Més tard, va ser utilitzat per estimar la línia de base, en connexió amb l'algoritme d'alineat "Parametric Time Warping" (Eilers, 2004), i va ser presentat amb més detalls a (Eilers Paul H. C., 2005). Recentment, una modificació de l'algoritme ALS anomenada *airPLS* es va presentar (Zhang et al., 2010), millorant els pesos del mètode ALS original.

El mètode ALS es basa en estimar una línia de base  $z$  donat un espectre  $y$ .  $z$  ha de ser més suau que  $y$ , però similar, de manera que ALS proposa minimitzar una funció cost inicialment donada per:

$$S = \sum_i d_i^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (\text{A.1})$$

on  $d_i = y_i - z_i$  són els residus de l'estimació i  $\Delta^2 z_i = z_i - 2z_{i-1} + z_{i-2}$ .

El primer terme de  $S$  és el terme de *fidelitat* de  $z$  a  $y$ , mentre que el segon terme imposa la *suavitat* de  $z$ . La suavitat es controla pel paràmetre  $\lambda$ , habitualment

en el rang  $10^2 \leq \lambda \leq 10^9$ . Aquesta funció de cost cal que es generalitzi introduint pesos  $w$ :

$$S = \sum_i w_i d_i^2 + \lambda \sum_i (\Delta^2 z_i)^2 \quad (\text{A.2})$$

Aquests pesos  $w$  s'introdueixen per tal de rebutjar penalitzacions a la funció de cost produïdes per regions on el senyal es troba per sobre la línia de base, és a dir regions amb pics. Així, el mètode ALS defineix  $w$  basant-se en un paràmetre  $p$  habitualment en el rang  $0.001 \leq p \leq 0.1$ . Així, s'aconsegueix que en les regions on no hi ha pics la línia de base travessi el soroll, mentre que en les regions on hi ha pics la línia de base es mantingui per sota el senyal.

$$w_i = \begin{cases} p & \text{if } d_i > 0 \\ 1 - p & \text{altrament} \end{cases} \quad (\text{A.3})$$

Com pot veure's, les regions on el senyal es troba per sobre la línia de base tindran una penalització molt més petita. Iterativament pot minimitzar-se  $S$ , trobant solucions convergents en menys de 20 iteracions habitualment.

Una limitació del mètode ALS es presenta en mostres de gran rang dinàmic, amb pics d'intensitats elevades. En aquest cas, l'estimació de la línia de base donada per ALS no convergirà a la línia de base real i o bé part de la línia de base penetrarà dins els pics més intensos o bé la línia de base s'ajustarà per sota el soroll i no travessant-lo.

La millora d'ALS anomenada airPLS i proposada a (Zhang et al., 2010), es fa amb dos objectius: Eliminar el paràmetre  $p$ , simplificant l'ús de l'algoritme i per millorar la qualitat de l'estimació fent servir pesos adaptats a la distància amb la línia de base segons:

$$w_i = \begin{cases} 0 & \text{if } d_i > 0 \\ \exp\left(\frac{-t \cdot |d_i|}{\sum_{d_i < 0} |d_i|}\right) & \text{altrament} \end{cases} \quad (\text{A.4})$$

on  $t$  és la iteració actual. Així, les regions on el senyal està per sobre la línia de base estimada són ignorades a la iteració següent.

Nosaltres proposem una modificació anomenada *psalsa* (algoritme ALS per senyals amb pics) amb una definició diferent per aquests pesos, molt més similar a la original de ALS:

$$w_i = \begin{cases} p \cdot e^{-\frac{d_i}{k}} & \text{if } d_i > 0 \\ 1 - p & \text{otherwise} \end{cases} \quad (\text{A.5})$$

La diferència amb el mètode original ALS és en els residus positius, on  $p$  es pondera per  $\exp\left(-\frac{d_i}{k}\right)$ . Les regions amb pics més elevats mostraran residus més grans, tenint pesos més petits, mentre que regions de soroll tindran residus petits i pesos semblants a  $p$ . Aquesta aproximació introdueix un nou paràmetre  $k$ , que controla com cauen els pesos. Aquest paràmetre es pot configurar a 5% de la intensitat màxima. Cal notar com si fem  $k \rightarrow \infty$  recuperem el mètode ALS original. En unes 5-10 iteracions *psalsa* és capaç de convergir correctament.

Una comparativa dels diferents mètodes d'estimació de línia de base es troba a A.3. També es mostra a la tesi com el valor del paràmetre  $k$  no és crític per l'estimació de la línia de base. A la figura A.4 es mostren uns cromatogrames de mostres reals, amb les línies de base estimades per cada mètode. Es pot veure com *psalsa* és capaç de travessar els pics creuant el soroll, i no passant-hi per sota.

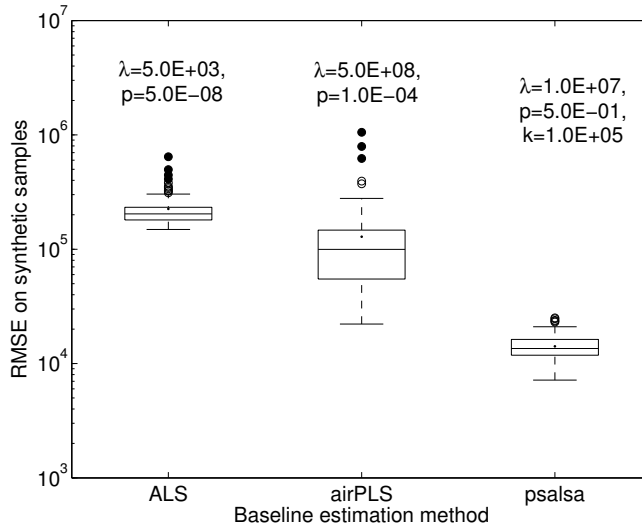


Figure A.3: Comparació de tres mètodes per cromatogrames sintètics

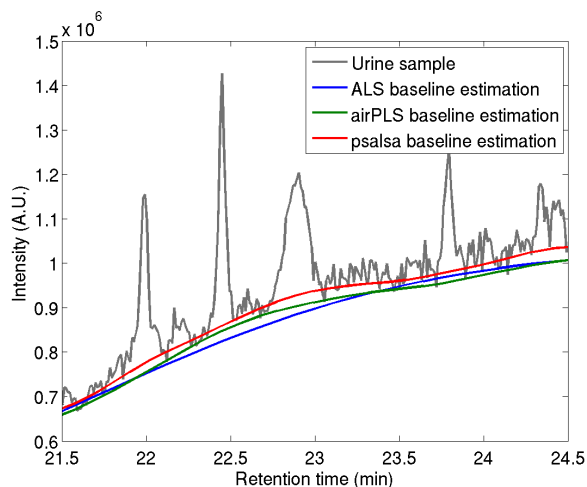


Figure A.4: Comparació de l'estimació de la línia de base en mostres reals

## A.2.2 Alineat

Variacions menors de la temperatura i pressió en una mesura poden afectar les mobilitats dels ions, que mouran les posicions dels pics a l'espectre IMS. Variacions en fluxos, temperatures i la degradació de la columna cromatogràfica poden fer moure també les posicions dels pics en temps de retenció. Per aquest motiu és necessari l'ús d'eines d'alineament que corregeixin aquestes derives.

En aquesta tesi hem fet una caracterització dels problemes d'alineament de mostres en un estudi al llarg de més de 10 mesos. Fent ús de calibrants externs hem vist la distribució dels desalineaments, resumida en les dispersions de la taula A.2.

Table A.2: Posicions originals dels pics

Nom	Cluster	Temps de deriva (ms)	Temps de retenció (s)
2-butanone	Dimer	$8.58 \pm 0.11$	$92 \pm 11$
2-pentanone	Dimer	$9.44 \pm 0.11$	$124 \pm 15$
2-hexanone	Dimer	$10.35 \pm 0.12$	$212 \pm 27$
2-heptanone	Dimer	$11.24 \pm 0.13$	$401 \pm 33$
2-heptanone	Monomer	$8.66 \pm 0.11$	$406 \pm 32$
2-octanone	Dimer	$12.12 \pm 0.13$	$584 \pm 36$



Table A.2: Posicions originals dels pics

Nom	Cluster	Temps de deriva (ms)	Temps de retenció (s)
2-octanone	Monomer	$9.15 \pm 0.11$	$590 \pm 36$
2-nonanone	Dimer	$12.96 \pm 0.13$	$1052 \pm 88$
2-nonanone	Monomer	$9.65 \pm 0.11$	$1053 \pm 87$

La correcció del temps de deriva s'ha fet mitjançant una correcció multiplicativa. Aquest és un mètode simple i lineal. Assumeix un canvi d'eix en temps de deriva de l'estil  $t' = kt$ , on  $k$  ha de ser estimat. La manera més habitual d'estimar  $k$  és fent servir el RIP, de manera que quedi ben alineat. Aquesta correcció és equivalent a la conversió de temps de deriva a mobilitats reduïdes.

Amb una correcció lineal dels temps de deriva, ajustada per tal que el RIP quedi alineat, ja es pot reduir en un ordre de magnitud els desalineaments en temps de deriva, com mostra la taula A.3.

Table A.3: Peak positions after drift time correction

Name	Cluster	Drift time (ms)	Drift time (corr.) (ms)
2-butanone	Dimer	$8.58 \pm 0.11$	$8.46 \pm 0.02$
2-pentanone	Dimer	$9.44 \pm 0.11$	$9.3 \pm 0.02$
2-hexanone	Dimer	$10.35 \pm 0.12$	$10.19 \pm 0.03$
2-heptanone	Dimer	$11.24 \pm 0.13$	$11.07 \pm 0.04$
2-heptanone	Monomer	$8.66 \pm 0.11$	$8.53 \pm 0.01$
2-octanone	Dimer	$12.12 \pm 0.13$	$11.94 \pm 0.05$
2-octanone	Monomer	$9.15 \pm 0.11$	$9.02 \pm 0.02$
2-nonanone	Dimer	$12.96 \pm 0.13$	$12.76 \pm 0.05$
2-nonanone	Monomer	$9.65 \pm 0.11$	$9.51 \pm 0.03$

Pel que fa el temps de retenció, hem fet servir splines cúbics monotònics per alinear-los. Fins a on podem saber, els splines cúbics monotònics no s'havien fet servir anteriorment per alinear dades de GC-IMS. Tanmateix, en aplicacions de cromatografia, la feina més similar va ser documentada per (Halang et al., 1978), on splines cúbics naturals s'utilitzen per l'alineat d'índexs de retenció. Més recentment, (Eilers, 2004), remarca les distorsions no uniformes dels temps de retenció observades a (Gong et al., 2004) en un instrument de cromatografia líquida d'alta pressió amb un detector de matriu de diodes (HPLC-DAD) i suggereix com una estratègia viable d'alineat l'ús de p-splines.

Els splines cúbics monotònics són adients perquè poden adaptar-se a variacions locals, tenen un comportament suau i preserven l'ordre d'elució (no esperem canvis en l'ordre d'elució en les mostres analitzades). En comparativa, els models lineals o polinomials no tenen la flexibilitat suficient per ajustar-se a les variacions del temps de retenció i la seva flexibilitat local és més limitada. Si bé és cert que l'ús de models lineals a trossos és possible, els models d'alineat a trossos no acostumen a estar ben comportats en els límits dels segments. Per exemple, el popular algoritme icoshift (Tomasi et al., 2011) ha de deixar valors buits a les vores dels segments per evitar que apareguin artefactes. Els splines cúbics monotònics, tal i com van descrits per (Hyman, 1983) compleixen totes les condicions descrites.

La taula A.4 mostra com queden alineats els temps de retenció fent una correcció lineal i una correcció mitjançant splines.

La correcció lineal té biaixos més elevats que la correcció amb splines perquè no té la flexibilitat suficient per copsar amb els canvis no-lineals.

Table A.4: Comparativa dels temps de retenció estimats

Pic de prova	Referència (s)	Linear (s)	Splines (s)
2-pentanone	95	$98 \pm 4$	$94 \pm 3$
2-hexanone	165	$178 \pm 8$	$169 \pm 5$
2-heptanone	343	$345 \pm 10$	$342 \pm 11$
2-octanone	527	$502 \pm 16$	$531 \pm 21$

Aquesta anàlisi ha contribuït a caracteritzar els desalineaments tant en temps de retenció com en temps de deriva. L'ús de splines cúbics monotònics per l'alineació del temps de retenció en mostres de GC-IMS ofereix una millora senzilla sobre les correccions lineals i desplaçaments habituals. A la tesi també se suggereix en base als resultats obtinguts un interval de dos dies per analitzar dues mostres de calibrants consecutives.

### A.3 Resolució multivariant de corbes en finestra mòbil (SW-MCR)

Les tècniques de separació cega de fonts (BSS) tenen per objectiu extreure de forma no supervisada un conjunt de senyals d'una mescla. Aquestes tècniques són útils en situacions amb una alta coelució, en les que la columna cromatogràfica és incapaç de separar completament els components de la mescla.

Una de les tècniques més esteses en l'àmbit de la quimiometria és la resolució multivariant de corbes amb mínims quadrats alternats (MCR-ALS) (Tauler, 1995). Aquesta tècnica es basa en fer una descomposició bilineal, on d'una banda tenim les concentracions de cada analit de la mostra i de l'altra l'espectre pur de l'analit corresponent, com es mostra a l'equació (A.6).

$$D = CS^T + E \quad (\text{A.6})$$

on:

- $D$  ( $M \times N$ ) és la matriu mesurada, amb un espectre per fila
- $C$  ( $M \times K$ ) són les abundàncies o concentracions de cada espectre per cada espectre pur
- $S$  ( $N \times K$ ) són els espectres purs
- $E$  és la matriu dels residus

Per tal de fer aquesta descomposició, MCR-ALS es basa en un procés iteratiu de mínims quadrats i en l'aplicació de restriccions de caire físic o químic per obtenir una descomposició interpretable, seguint els passos següents:

1. Eliminar soroll de  $D$ . Per eliminar el soroll es fa una descomposició PCA amb  $K$  components principals i es reconstrueix la matriu a partir dels scores i els loadings del PCA.
2. Estimar els perfils de concentració per mínims quadrats:

$$C = \operatorname{argmin}_C \|D * -CS^T\|^2 \quad (\text{A.7})$$

3. Imposar restriccions sobre els perfils de concentració
4. Estimar els espectres purs fent servir mínims quadrats:

$$S = \operatorname{argmin}_S \|D * -CS^T\|^2 \quad (\text{A.8})$$

5. Imposar restriccions sobre els espectres purs
6. Iterar els passos 2-5 fins a convergir

La clau doncs per obtenir bons perfils de concentració i espectres purs depèn en l'estimació del nombre de components de la mescla, la inicialització dels espectres purs i la imposició de restriccions que assegurin solucions física i químicament raonables. Detalls sobre els mètodes per inicialitzar i les restriccions més habituals es donen a la tesi.

MCR-ALS es basa en una minimització per mínims quadrats de l'error global de la factorització. Com mostren els resultats de la tesi, els pics locals amb intensitats més baixes que apareixen en regions amb més co-elució no són detectats per MCR-ALS correctament, a causa que comprenen una contribució a l'error comparable o fins i tot inferior que el soroll global de la mostra. En aquests casos, incrementar el nombre de components de la mostra porta a extreure també compostos “espuris”, no desitjats i sense significat físic en comptes d'aquests compostos locals.

Per tal de superar la limitació descrita, proposem aplicar MCR-ALS en finestres petites i parcialment solapades, fent llesques de la matriu de la mostra en l'eix de temps de retenció. A més, les finestres es fan parcialment solapades per evitar partir els pics a les vores de la finestra, i per rebutjar els compostos espuris que puguin aparèixer esporàdicament en alguna finestra.

Primer, les estimacions inicials d'espectres purs i perfils de concentració s'obtenen a cada finestra. El nombre de components per cada finestra s'estima fent servir un llindar en els valors singulars. Donades les estimacions inicials, apliquem MCR-ALS a cada finestra, i n'extraïem un conjunt de perfils de concentració i espectres purs per cada finestra.

Finalment, els resultats de totes les finestres es combinen en un únic conjunt de perfils de concentració i espectres purs, representatius de tota la mostra. Per fer-ho, els compostos es segueixen al llarg de finestres consecutives, en base a una figura de similitud d'espectres. L'angle entre dos espectres purs  $s_i$  i  $s_j$  es calcula segons (A.9).

$$\theta_{i,j} = \arccos \left( \frac{s_i s_j}{\|s_i\| \|s_j\|} \right) \quad (\text{A.9})$$

La figure A.5 mostra un diagrama amb un exemple de quatre compostos detectats en tres finestres. L'enllaç entre dos espectres de finestres consecutives es forma només si el seu angle es troba per sota un llindar. En aquesta figura, els compostos C1 i C2 se segueixen al llarg de les finestres N a N+2, mentre que el compost C3 desapareix a la finestra N+1 perquè no s'hi pot establir cap enllaç amb la darrera finestra. El compost C4 no apareix fins a la finestra N+1. El darrer espectre de la finestra N+1 no estableix cap enllaç amb cap altra finestra, i per tant es considera espuri i es descarta del conjunt de compostos finals.

La metodologia s'ha aplicat a mostres d'oli d'oliva mesurades amb MCC-IMS, imposant les restriccions de no-negativitat a espectres i concentracions, sistema tancat en els perfils de concentració i unimodalitat dels espectres purs. L'angle entre els compostos ha de ser inferior a 15 graus per tal que es considerin el

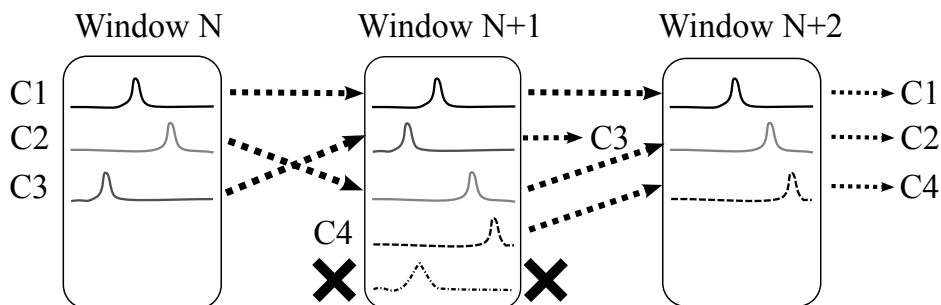


Figure A.5: Diagrama del seguiment dels espectres al llarg de tres finestres. Els enllaços entre espectres s'estableixen si el seu angle es troba per sota d'un llindar.

mateix compost. Aplicant aquesta metodologia, SW-MCR és capaç d'extreure més compostos, com es mostra a la taula A.5.

Així, hem presentat una tècnica nova per la millora de la resolució multivariant de corbes de mostres de GC-IMS, mostrant que la deconvolució cega de pics es pot aplicar amb èxit a aquest instrument analític per tal d'extreure característiques i resoldre coelucions fins i tot quan l'ús convencional de MCR-ALS és incapaç de discriminar les intensitats més baixes del soroll de la mostra.

## A.4 Aplicacions

### A.4.1 Anàlisi de qualitat d'oli d'oliva mitjançant GC-IMS

La primera aplicació està relacionada amb la prevenció de frau en alimentació. Diferents tipus d'oli d'oliva tenen característiques organolèptiques diferents i per tant diferent preu de mercat. Una classificació simple i acurada dels tipus d'oli d'oliva és desitjable, en tant que la indústria de l'oli pot reduir despeses i prevenir el frau. L'oli d'oliva pot categoritzar-se en tres grups d'acord amb la seva qualitat: "Verge extra", "Verge" i "Llampant".

En aquesta aplicació utilitzem algunes de les tècniques de processat de dades descrites anteriorment, combinades amb PLS-DA, un classificador lineal, i doble validació creuada. La metodologia pel modelat de les dades ve descrita a la figura A.6.

Un resum dels resultats queda plasmat a la figura A.7. A la part esquerra pot apreciar-se les mostres d'entrenament i de test a l'espai format per les dues

Table A.5: Localització de 22 pics triats aleatòriament de la mostra, en la deconvolució MCR-ALS i en la deconvolució SW-MCR. Els rangs dels temps de retenció són consistents amb els pics detectats, tot i que MCR-ALS no ha estat capaç d'extreure 9/22 pics.

# Pic	Temps de deriva (ms)	Int. màx. (a.u.)	Rang de temps de retenció (s)		
			Mostra	MCR-ALS	SW-MCR
1	6.45 (RIP)	3951	Sempre	Sempre	Sempre
2	7.30	3936	1-4	0-10	0-4
3	7.60	1002	3-7	2-7	3-9
4	7.75	400	4-10	NF	4-9
5	8.30	782	5-9	4-8	4-10
6	9.10	177	5-12	NF	4-10
7	8.10	695	4-9	0-20	4-10
8	8.78	731	5-12	4-12	5-12
9	8.60	623	4-12	4-7	5-13
10	7.15	474	4-13	3-7	5-15
11	8.90	2100	8-13	8-12	8-13
12	8.10	425	11-19	NF	12-20
13	6.75	490	11-23	NF	12-23
14	10.30	1190	20-27	22-27	21-27
15	7.20	188	21-29	NF	21-28
16	8.20	481	21-32	NF	21-31
17	8.50	390	30-38	30-35	30-37
18	7.80	317	28-40	NF	30-40
19	7.30	445	31-40	NF	32-40
20	9.90	2200	40-55	40-60	40-50
21	8.70	220	50-63	NF	50-56
22	7.65	650	45-80	50-80	55-80

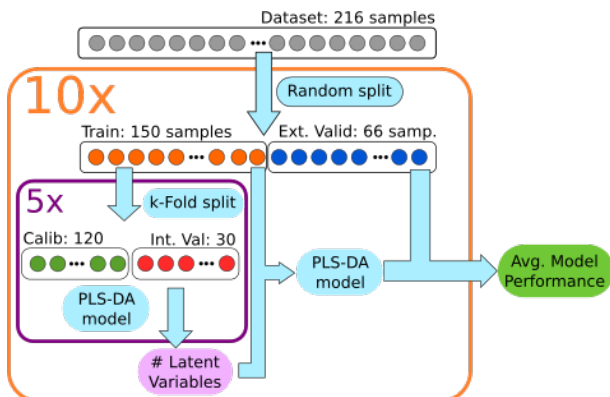


Figure A.6: Doble validació creuada per avaluar la classificació del model. Divisió en entrenament i test, validació interna tipus k-fold,  $k = 5$ . El procés va repetir-se 10 vegades per poder obtenir una estimació de la incertesa de la taxa d'encert del model.

primeres variables latents del model PLS-DA. La separació entre les categories d'oli verge extra i llampant és força clara, mentre que la d'oliva verge mostra certa confusió. Tanmateix, les variables latents següents (no representades) milloren la separació. A la part dreta de la figura es mostra el loading corresponent a la primera variable latent. Les mostres amb valors més elevats que la mitjana en les zones liles tindran un score més elevat a la primera variable latent i per tant segons els scores serà més probable que siguin mostres d'oli verge extra. Els models, entrenats amb 6 variables latents, són avaluats amb les mostres de validació externa, resultant en una taxa d'encert del  $85\% \pm 5\%$ .

#### A.4.2 Cerca de biomarcadors de càncer de pròstata en els volàtils de la orina

La segona aplicació s'enfoca en l'anàlisi dels volàtils en mostres d'orina humana fent servir cromatografia de gasos espectrometria de masses, amb l'objectiu de detectar biomarcadors capaços de discriminar individus que pateixen càncer de pròstata.

El diagnòstic del càncer de pròstata es confirma avui dia mitjançant una biòpsia. Les biòpsies són invasives i incòmodes pels pacients, a més de suposar un cost significatiu. Acostumen a fer-se després d'un resultat positiu en el test sanguini PSA (Prostate Specific Antigen) que actua com a filtre previ. Tanmateix, PSA

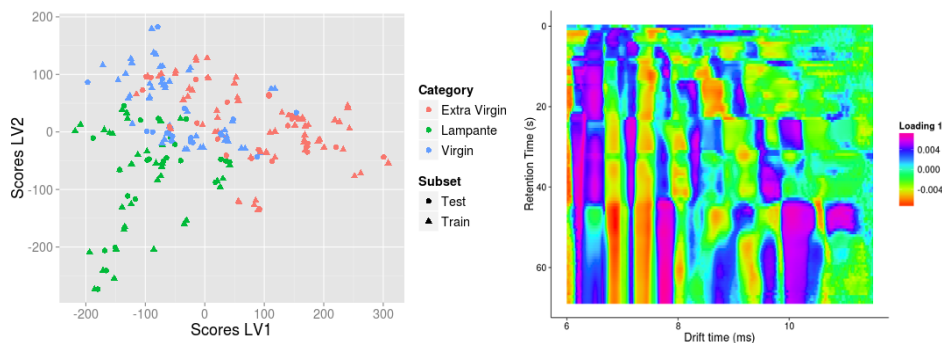


Figure A.7: Scores per les dues primeres variables latents del model i loadings de la primera variable latent.

és un test amb controvèrsia a causa de l'elevat nombre de falsos positius que dona, principalment relacionats amb condicions no cancerígenes com la hiperplàsia benigna de pròstata (Thompson et al., 2004). Tenint un resultat positiu en només el 30% de les biòpsies fetes indica que el 70% de biòpsies que es fan són innecessàries, i que per tant trobar altres biomarcadors no invasius i específics del càncer de pròstata reduiria el nombre de biòpsies fetes, estalviant incomoditats als pacients, temps i recursos.

Estudis recents confirmen progressos en la detecció de càncer de pròstata fent servir gossos que oloren orina (Cornu et al., 2011). A (Khalid et al., 2015), els autors informe de la detecció de potencials biomarcadors, amb una classificació, d'entre el 60% i el 70%, no gaire superior al PSA.

Després de fer l'anàlisi de mostres d'orina de pacients, la línia de base és corregida fent ús de *psalsa*, se'n redueix el soroll i s'extreuen els pics mitjançant una deconvolució de Bieller - Biemann. El resultat de la detecció de pics es mostra a la figura A.8. Després d'aplicar l'algorisme descrit a (Robinson et al., 2007) per agrupar els pics corresponents al mateix analit de diferents mostres, vam aplicar un test de Wilcoxon sobre la taula de pics, sense obtenir diferències significatives en cap dels compostos detectats. La figura A.9 mostra els pics amb diferències més grans en les medianes, i es pot veure com no hi ha diferències clares.

Després de validar que l'error no es troba en la metodologia d'anàlisi (repetint l'anàlisi amb altres tècniques, veure (Macías, 2017)), entenem que el problema ha de ser o a la nostra instrumentació, o al protocol experimental o en la reproduïbilitat de les anàlisi. Més enllà d'això, aquesta aplicació mostra algunes de les eines descrites a la tesi, i és una mostra de com difícil pot ser obtenir resultats reproduïbles en aquest camp.



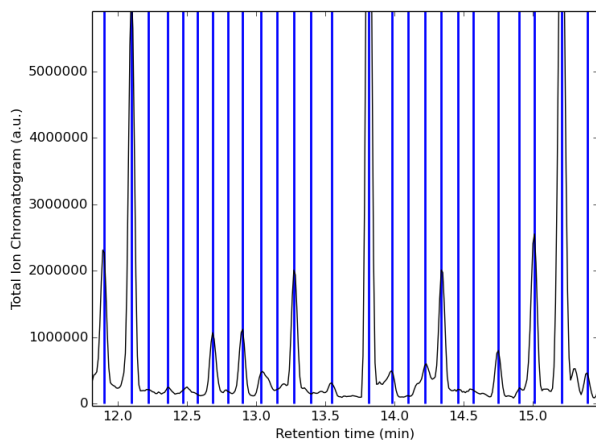


Figure A.8: Detecció de pics a mostres GC-MS.

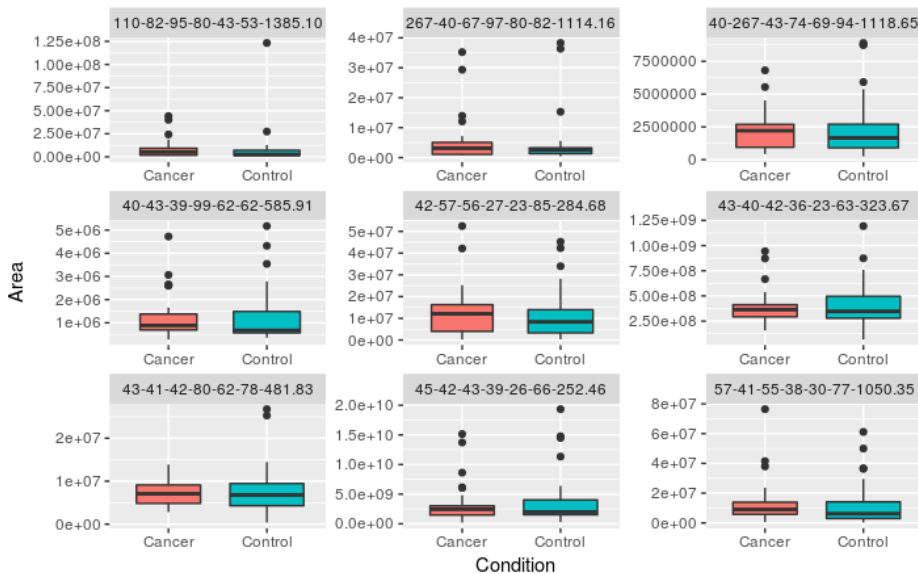


Figure A.9: Boxplot dels pics amb diferències més grans en les medianes de les àrees per pacients de càncer i control.

# Appendix B

## Conclusions

Aquesta tesi estudia eines i algorismes per l'anàlisi de dades d'instrumentació química acoblada, enfocant-se en cromatografia de gasos - espectrometria de mobilitat d'ions.

- L'estudi a la secció 2.2 de diverses tècniques d'estimació de la línia de base per mostres de GC-MS i GC-IMS ha mostrat una limitació de la tècnica de mínims quadrats asimètrics (ALS) per aquelles mostres que tenen pics intensos o un rang dinàmic gran. La modificació dels pesos de l'algoritme ALS per tal que depenguin de la magnitud del residu ha mostrat ser efectiva per superar la limitació descrita, millorant l'estimació de la línia de base tant en mostres sintètiques com reals.
- La caracterització dels desalineaments de la secció 2.3 elaborada a un conjunt de dades de GC-IMS que s'estén més de 10 mesos ha mostrat que les correccions a un primer ordre del temps de deriva poden corregir-se amb una transformació lineal tenint un factor de correcció inferior al 3%. A més, aquest factor de correcció com es mostra a la figura 2.14 està agrupat pel dia de l'anàlisi, indicant una correlació de desalineaments intra-dia en el temps de deriva. Tanmateix, els desalineaments en temps de retenció es beneficien més de correccions no lineals, de manera que hem explorat com els splines cúbics monotònics milloren les correccions lineals gràcies a la seva flexibilitat. Basant-nos en aquests resultats hem suggerit un temps òptim entre dues mesures consecutives de calibrants de dos dies. Tot i la correcció proposada, basada en calibrants externs, creiem que l'ús de calibrants interns encara és convenient per la correcció dels temps de retenció, per tal de tenir més seguretat en l'alineament i cobrir les variacions no sistemàtiques.

- Una nova tècnica per millorar la resolució quimiomètrica de mostres de GC-IMS ha estat presentada mostrant per primer cop que aquestes tècniques són també adequades per aquesta instrumentació.
- La tècnica SW-MCR ha estat provada en mostres d'oli d'oliva, que presenten una important coelució de components, tal i com es mostra en l'amplitud dels pics del RIP invers. Els pics en coelució no es resolen correctament utilitzant mètodes MCR-ALS convencionals, ja que els pics de menor intensitat no poden ser discriminats del soroll de la mostra. A més, els resultats de MCR-ALS requereixen supervisió per l'aparició de compostos espuris.
- Per contra, utilitzant el mètode SW-MCR proposat, hem estat capaços de deconvolucionar els espectres purs i els perfils de concentració de la majoria de pics, fins i tot els d'intensitats menors, rebutjant pics espuris de manera automàtica.
- Hem mostrat l'aplicació d'aquestes eines a la classificació d'olis d'oliva d'acord amb la seva qualitat entrenant un classificador i utilitzant metodologies de validació recomanades.
- També hem aplicat alguns dels mètodes treballats a mostres de GC-MS, buscant biomarcadors de càncer en volàtils d'orina, identificant compostos que tot i tenir potencial de ser biomarcadors segons la literatura, no han donat bones capacitats de predicció.

Esperem que la implementació d'aquestes tècniques en un repositori obert faciliti la seva adopció i la comparativa per altres investigadors.

# Appendix C

## Publications

### C.1 Contributions to open source packages

During the development of this thesis, I have put effort in improving the ecosystem of data analysis tools, fixing bugs and offering feature improvements to widespread R packages like `readr`, `dplyr`, `readxl`, `xlsx`, `plotly`, `scales` or `broom` that belong to the Top 100 most downloaded packages in CRAN, the Comprehensive R Archive Network. I have offered fixes and improvements to other popular packages like `rmarkdown`, `htmlTable`, or `shinyFiles` and I have submitted my own R package `condformat`.

In python, my major contribution has been a package to make parallelization easier, that so far has been used in astrophysics, for simulations in the dynamics of spinning black-hole binaries (Gerosa and Kesden, 2016), in deep learning, to train generative adversarial networks (Isola et al., 2016) and in epigenetics on research related to chronic lymphocytic leukaemia (Rendeiro et al., 2016).

### C.2 Publications

- S. Oller-Moreno, O. Cominetti, A. Núñez Galindo, I. Irincheeva, J. Corthésy, A. Astrup, W. H.M. Saris, J. Hager, M. Kussmann, L. Dayon *The differential plasma proteome of obese and overweight individuals undergoing a nutritional weight loss and maintenance intervention* Proteomics: Clinical Applications 2017 DOI: 10.1002/prca.201600150

- O. Cominetti, A. Núñez Galindo, J. Corthésy, S. Oller-Moreno, I. Irincheeva, A. Valsesia, A. Astrup, W. Saris, J. Hager, M. Kussmann, L. Dayon *Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry* J. Proteome Res, 2016 DOI: 10.1021/acs.jproteome.5b00901
- S. Oller-Moreno, G.Singla-Buxarrais, J.M. Jiménez-Soto, A.Pardo, R.Garrido-Delgado, L.Arce, S.Marco *Sliding window multi-curve resolution: Application to gas chromatography-ion mobility spectrometry* Sensors and Actuators B: Chemical, 2015 DOI: 10.1016/j.snb.2015.02.108

### C.3 Oral Presentations in conferences

- Oller-Moreno, M. Padilla, J.M. Jiménez-Soto, A. Pardo, S. Marco *Graphical User Interface for IMS and GC-IMS data preprocessing. Example of application*, 25th Intl. Conf. on Ion Mobility Spectrometry, Boston, 2016
- S. Oller-Moreno, J. Fonollosa, J.M. Jiménez-Soto, R. Garrido-Delgado, L. Arce, A. Pardo, S. Marco *Spectra Alignment Characterization and Correction in Gas-Chromatography – Ion Mobility Spectrometry*, XVI Chemometrics in Analytical Chemistry Conference, Barcelona 2016
- S. Oller-Moreno, S. Rica, J.M. Jiménez-Soto, J. Xaubet, A. Pardo, S.Marco, *Toolbox for MCC-IMS and IMS data analysis*, Intl. Conf. on Ion Mobility Spectrometry, Córdoba 2015.
- S. Oller-Moreno, A. Pardo, JM Jiménez-Soto, J. Samitier, S. Marco *Adaptive Asymmetric Least Squares baseline estimation for analytical instruments* Proc. Intl. Conf. on Communication and Signal Processing, Castelldefels 2014

### C.4 Posters

- S. Marco, G.Singla-Buxarrais, S.Oller-Moreno, JM Jiménez-Soto, A. Pardo, R.Delgado-Garrido, L.Arce *Chemometric Resolution for hyphenated Gas Chromatography Ion Mobility Spectrometry*, Proc. 15th Intl. Meeting on Chemical Sensors Buenos Aires, 2014.
- S.Oller-Moreno, A.Pardo, S.Marco, J.Samitier *Preprocessing techniques for GC-MS metabolomics data*, Proc. 6th IBEC Symposium, 2013
- S. Oller-Moreno, R.Garrido-Delgado, L. Arce, M. Valcárcel, A.Pardo, S.Marco *Multivariate Curve Resolution - Alternating Least Squares applied to GC-IMS olive oil measurements*, Proc. XIII - Chemometrics in Analytical Chemistry, Budapest 2012

# Bibliography

- Alonso, R., Rodríguez-Estévez, V., Domínguez-Vidal, A., Ayora-Cañada, M. J., Arce, L., and Valcárcel, M. (2008). Ion mobility spectrometry of volatile compounds from Iberian pig fat for fast feeding regime authentication. *Talanta*, 76(3):591–6, ISSN: 1873–3573, DOI: 10.1016/j.talanta.2008.03.052, <http://www.sciencedirect.com/science/article/pii/S0039914008002476>.
- Armenta, S., Alcalá, M., and Blanco, M. (2011). A review of recent, unconventional applications of ion mobility spectrometry (IMS). *Analytica Chimica Acta*, 703(2):114–123, ISSN: 00032670, DOI: 10.1016/j.aca.2011.07.021, <http://linkinghub.elsevier.com/retrieve/pii/S0003267011009627>.
- Bader, S., Urfer, W., and Baumbach, J. I. (2008). Preprocessing of ion mobility spectra by lognormal detailing and wavelet transform. *International Journal for Ion Mobility Spectrometry*, 11(1-4):43–49, ISSN: 1435–6163, DOI: 10.1007/s12127-008-0005-6, <http://link.springer.com/10.1007/s12127-008-0005-6>.
- Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67(339):687–690, ISBN: Journal of the American Statistical Association, Vol. 67, No. 339, September 1972: pp. 687–690, ISSN: 1537274X, DOI: 10.1080/01621459.1972.10481279.
- Baumbach, J. I. (2009). Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *Journal of breath research*, 3(3):034001, ISSN: 1752-7163, DOI: 10.1088/1752-7155/3/3/034001, [http://iopscience.iop.org/1752-7163/3/3/034001/pdf/1752-7163\\_3\\_3\\_034001.pdf](http://iopscience.iop.org/1752-7163/3/3/034001/pdf/1752-7163_3_3_034001.pdf).
- Baumbach, J. I., Berger, D., Leonhardt, J. W., and Klockow, D. (1993). Ion Mobility Sensor In Environmental Analytical Chemistry—Concept And First Results. *International Journal of Environmental Analytical Chemistry*, 52(1-4):189–193, ISSN: 0306-7319, DOI: 10.1080/03067319308042859, <http://www.tandfonline.com/doi/abs/10.1080/03067319308042859>.

- Baumbach, J. I., Sielemann, S., Xie, Z., and Schmidt, H. (2003). Detection of the Gasoline Components Methyl tert -Butyl Ether, Benzene, Toluene, and m -Xylene Using Ion Mobility Spectrometers with a Radioactive and UV Ionization Source. *Analytical Chemistry*, 75(6):1483–1490, ISSN: 0003-2700, DOI: 10.1021/ac020342i, <http://dx.doi.org/10.1021/ac020342i>.
- Begley, C. G. and Ioannidis, J. P. A. (2015). Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circulation Research*, 116(1):116–126, ISSN: 0009-7330, DOI: 10.1161/CIRCRESAHA.114.303819, <http://circres.ahajournals.org/cgi/doi/10.1161/CIRCRESAHA.114.303819>.
- Berant, Z., Karpas, Z., and Shahal, O. (1989). Effects of temperature and clustering on mobility of ions in carbon dioxide. *The Journal of Physical Chemistry*, 93(21):7529–7532, ISSN: 0022-3654, DOI: 10.1021/j100358a052, <http://pubs.acs.org/doi/abs/10.1021/j100358a052>.
- Bermejo, S., Jutten, C., and Cabestany, J. (2006). ISFET source separation: Foundations and techniques. *Sensors and Actuators B: Chemical*, 113(1):222–233, ISSN: 09254005, DOI: 10.1016/j.snb.2005.02.050, <http://www.sciencedirect.com/science/article/pii/S0925400505002352>.
- Bloemberg, T. G., Gerretzen, J., Wouters, H. J., Gloerich, J., van Dael, M., Wessels, H. J., van den Heuvel, L. P., Eilers, P. H., Buydens, L. M., and Wehrens, R. (2010). Improved parametric time warping for proteomics. *Chemometrics and Intelligent Laboratory Systems*, 104(1):65–74, ISSN: 01697439, DOI: 10.1016/j.chemo1ab.2010.04.008, <http://www.sciencedirect.com/science/article/pii/S0169743910000572>.
- Bödeker, B., Vautz, W., and Baumbach, J. I. (2008). Peak comparison in MCC/IMS-data—searching for potential biomarkers in human breath data. *International Journal for Ion Mobility Spectrometry*, 11(1-4):89–93, ISSN: 1435-6163, DOI: 10.1007/s12127-008-0013-6, <http://link.springer.com/10.1007/s12127-008-0013-6>.
- Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3):83–92, ISSN: 0014-5793, DOI: 10.1016/j.febslet.2004.07.055, <http://www.ncbi.nlm.nih.gov/pubmed/15327980>.
- Broadhurst, D. I. and Kell, D. B. (2007). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171–196, ISSN: 1573-3882, DOI: 10.1007/s11306-006-0037-z, <http://link.springer.com/10.1007/s11306-006-0037-z>.

- Budde, K. J. (1995). Application of Ion Mobility Spectrometry to Semiconductor Technology: Outgassings of Advanced Polymers under Thermal Stress. *Journal of The Electrochemical Society*, 142(3):888, ISSN: 00134651, DOI: 10.1149/1.2048553, <http://jes.ecsdl.org/content/142/3/888.abstract>.
- Bunkowski, A. (2012). *MCC-IMS data analysis using automated spectra processing and explorative visualisation methods*. PhD thesis, Bielefeld University, <http://pub.uni-bielefeld.de/publication/2517237>.
- Buszewski, B., Ulanowska, A., Kowalkowski, T., and Cieliski, K. (2012). Investigation of lung cancer biomarkers by hyphenated separation techniques and chemometrics. *Clinical Chemistry and Laboratory Medicine*, 50(3):573–581, ISBN: 1434-6621, ISSN: 14346621, DOI: 10.1515/cc1m.2011.769.
- Buxton, T. L. and Harrington, P. d. B. (2001). Rapid multivariate curve resolution applied to identification of explosives by ion mobility spectrometry. *Analytica Chimica Acta*, 434(2):269–282, ISSN: 00032670, DOI: 10.1016/S0003-2670(01)00839-X, <http://www.sciencedirect.com/science/article/pii/S000326700100839X>.
- Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, ISSN: 00189219, DOI: 10.1109/5.720250, <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=720250>.
- Carteret, C., Dandeu, A., Moussaoui, S., Muhr, H., Humbert, B., and Plasari, E. (2009). Polymorphism Studied by Lattice Phonon Raman Spectroscopy and Statistical Mixture Analysis Method. Application to Calcium Carbonate Polymorphs during Batch Crystallization. *Crystal Growth & Design*, 9(2):807–812, ISSN: 1528-7483, DOI: 10.1021/cg800368u, <http://dx.doi.org/10.1021/cg800368u>.
- Cichocki, A. and Amari, S.-i. (2002). *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, Volum 1*. John Wiley & Sons, Ltd, ISBN: 0471607916, [http://books.google.com/books?hl=ca&lr=&id=IHp0q\\_TNGEkC&pgis=1](http://books.google.com/books?hl=ca&lr=&id=IHp0q_TNGEkC&pgis=1).
- Cichocki, A., Zdunek, R., and Amari, S. (2006). New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V-621–V-624. IEEE, ISBN: 1-4244-0469-X, ISSN: 1520-6149, DOI: 10.1109/ICASSP.2006.1661352, <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1661352>.



- Cline, J. and Hobbs, J. (1972). Laboratory Evaluation of Detectors of Explosives' Effluents. Technical report, Department of Transportation, USA, Cambridge, MA, <http://ntl.bts.gov/lib/46000/46500/46575/DOT-TSC-OST-72-27.pdf>.
- Cominetti, O., Núñez Galindo, A., Corthésy, J., Oller Moreno, S., Irincheeva, I., Valsesia, A., Astrup, A., Saris, W., Hager, J., Kussmann, M., and Dayon, L. (2016). Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry. *Journal of Proteome Research*, 15(2), ISSN: 15353907 15353893, DOI: 10.1021/acs.jproteome.5b00901.
- Commission, E. U. and others (1991). Regulation EEC/2568/91 on the characteristics of olive and olive pomace oils and their analytical methods. *Off. J. Eur. Communities L*, 248:1–83, <http://data.europa.eu/eli/reg/1991/2568/oj>.
- Conte, L. S., Moret, S., Bortolomeazzi, R., and Pizzale, L. (1999). The advancement of the assessment of food quality control as stressed by recent developments in analytical chemistry. *Annali di Chimica*, 89(9-10):805–816, ISSN: 00034592, <http://www.scopus.com/inward/record.url?eid=2-s2.0-0039623036&partnerID=tZOtx3y1>.
- Cornu, J.-N., Cancel-Tassin, G., Ondet, V., Girardet, C., and Cussenot, O. (2011). Olfactory Detection of Prostate Cancer by Dogs Sniffing Urine: A Step Forward in Early Diagnosis. *European Urology*, 59(2):197–201, ISSN: 03022838, DOI: 10.1016/j.eururo.2010.10.006, <http://linkinghub.elsevier.com/retrieve/pii/S0302283810009449>.
- Cumeras, R., Gràcia, I., Figueras, E., Fonseca, L., Santander, J., Salleras, M., Calaza, C., Sabaté, N., and Cané, C. (2012). Finite-element analysis of a miniaturized ion mobility spectrometer for security applications. *Sensors and Actuators B: Chemical*, 170:13–20, ISSN: 09254005, DOI: 10.1016/j.snb.2010.11.047, <http://www.sciencedirect.com/science/article/pii/S0925400510009160>.
- de Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., Osborne, D., and Ratcliffe, N. M. (2014). A review of the volatiles from the healthy human body. *Journal of Breath Research*, 8(1):014001, ISSN: 1752-7155, DOI: 10.1088/1752-7155/8/1/014001, <http://stacks.iop.org/1752-7163/8/i=1/a=014001?key=crossref.b6b6a6911efb2c74c533c2e3a6bae189>.
- Diewok, J., de Juan, A., Maeder, M., Tauler, R., and Lendl, B. (2003). Application of a Combination of Hard and Soft Modeling for Equilibrium Systems to the Quantitative Analysis of pH-Modulated Mixture Samples. *Analytical Chemistry*, 75(3):641–647, ISSN: 0003-2700, DOI: 10.1021/ac026248j, <http://dx.doi.org/10.1021/ac026248j>.

- Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, ISBN: 0006-3444, ISSN: 00063444, DOI: 10.1093/biomet/81.3.425.
- Duarte, L. T., Moussaoui, S., and Jutten, C. (2014). Source Separation in Chemical Analysis : Recent achievements and perspectives. *IEEE Signal Processing Magazine*, 31(3):135–146, ISSN: 1053-5888, DOI: 10.1109/MSP.2013.2296099, <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6784054>.
- Eatherton, R. L., Siems, W. F., and Hill, H. H. (1986). Fourier transform ion mobility spectrometry of barbiturates after capillary gas chromatography. *Journal of High Resolution Chromatography*, 9(1):44–48, ISSN: 0935-6304, DOI: 10.1002/jhrc.1240090108, <http://doi.wiley.com/10.1002/jhrc.1240090108>.
- Eiceman, G., Karpas, Z., Hill, H. H., and Jr. (2013). *Ion Mobility Spectrometry, Third Edition*. CRC Press, ISBN: 1439859973, <http://books.google.com/books?hl=en&lr=&id=tnlcAgAAQBAJ&pgis=1>.
- Eiceman, G. A. and Feng, Y. (2009). Limits of separation of a multi-capillary column with mixtures of volatile organic compounds for a flame ionization detector and a differential mobility detector. *Journal of Chromatography A*, 1216(6):985–993, ISSN: 00219673, DOI: 10.1016/j.chroma.2008.11.091.
- Eiceman, G. A., Yuan-Feng, W., Garcia-Gonzalez, L., Harden, C. S., and Shoff, D. B. (1995). Enhanced selectivity in ion mobility spectrometry analysis of complex mixtures by alternate reagent gas chemistry. *Analytica Chimica Acta*, 306(1):21–33, ISSN: 00032670, DOI: 10.1016/0003-2670(94)00668-C, <http://www.sciencedirect.com/science/article/pii/000326709400668C>.
- Eilers, P. H. C. (2003). A Perfect Smoother. *Analytical Chemistry*, 75(14):3631–3636, ISSN: 0003-2700, DOI: 10.1021/ac034173t, <http://pubs.acs.org/doi/abs/10.1021/ac034173t>.
- Eilers, P. H. C. (2004). Parametric Time Warping. *Analytical Chemistry*, 76(2):404–411, ISSN: 0003-2700, DOI: 10.1021/ac034800e, <http://pubs.acs.org/doi/abs/10.1021/ac034800e>.
- Eilers Paul H. C., B. H. F. M. (2005). Baseline Correction with Asymmetric Least Squares Smoothing. Technical report, Leiden University Medical Centre, Leiden.
- Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., and Buydens, L. M. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, ISSN:

- 01659936, DOI: 10.1016/j.trac.2013.04.015, <http://linkinghub.elsevier.com/retrieve/pii/S0165993613001465>.
- Ewing, R. (2001). A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds. *Talanta*, 54(3):515–529, ISSN: 00399140, DOI: 10.1016/S0039-9140(00)00565-8, <http://www.sciencedirect.com/science/article/pii/S0039914000005658>.
- Felinger, A. (1998). *Data Analysis and Signal Processing in Chromatography*. Elsevier, ISBN: 9780080525563.
- Fuchs, P., Loeseken, C., Schubert, J. K., and Miekisch, W. (2010). Breath gas aldehydes as biomarkers of lung cancer. *International Journal of Cancer*, 126(11):2663–2670, ISBN: 0020-7136, ISSN: 00207136, DOI: 10.1002/ijc.24970.
- Gan, F., Ruan, G., and Mo, J. (2006). Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2 SPEC. ISS):59–65, ISBN: 0169-7439, ISSN: 01697439, DOI: 10.1016/j.chemolab.2005.08.009.
- Garrido-Delgado, R., Arce, L., and Valcárcel, M. (2012). Multi-capillary column-ion mobility spectrometry: a potential screening system to differentiate virgin olive oils. *Analytical and bioanalytical chemistry*, 402(1):489–98, ISSN: 1618-2650, DOI: 10.1007/s00216-011-5328-1, <http://www.ncbi.nlm.nih.gov/pubmed/21866400>.
- Garrido-Delgado, R., Dobao-Prieto, M. d. M., Arce, L., and Valcárcel, M. (2015). Determination of volatile compounds by GC-IMS to assign the quality of virgin olive oil. *Food Chemistry*, 187:572–579, ISSN: 03088146, DOI: 10.1016/j.foodchem.2015.04.082, <http://linkinghub.elsevier.com/retrieve/pii/S0308814615006287>.
- Geladi, P. and Hopke, P. K. (2008). Editorial: Is there a future for chemometrics? Are we still needed? *Journal of Chemometrics*, 22(5):289–290, ISSN: 08869383, DOI: 10.1002/cem.1141, <http://doi.wiley.com/10.1002/cem.1141>.
- Gerosa, D. and Kesden, M. (2016). PRECESSION: Dynamics of spinning black-hole binaries with python. DOI: 10.1103/PhysRevD.93.124066, <http://arxiv.org/abs/1605.01067><http://dx.doi.org/10.1103/PhysRevD.93.124066>.
- Gong, F., Liang, Y.-Z., Fung, Y.-S., and Chau, F.-T. (2004). Correction of retention time shifts for chromatographic fingerprints of herbal medicines. *Journal of Chromatography A*, 1029(1-2):173–183, ISSN:

- 00219673, DOI: 10.1016/j.chroma.2003.12.049, <http://linkinghub.elsevier.com/retrieve/pii/S0021967303023616>.
- Gourvéneç, S., Massart, D., and Rutledge, D. (2002). Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the Orthogonal Projection Approach and in the SIMPLEx Interactive Self-modelling Mixture Analysis approach. *Chemometrics and Intelligent Laboratory Systems*, 61(1-2):51–61, ISSN: 01697439, DOI: 10.1016/S0169-7439(01)00172-1, <http://www.sciencedirect.com/science/article/pii/S0169743901001721>.
- Griffin, J. (2003). Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Current Opinion in Chemical Biology*, 7(5):648–654, ISSN: 13675931, DOI: 10.1016/j.cbpa.2003.08.008, <http://linkinghub.elsevier.com/retrieve/pii/S136759310300111X>.
- Guamán Novillo, A. V. (2015). *Multivariate Signal Processing for Quantitative and Qualitative Analysis of Ion Mobility Spectrometry data, applied to Biomedical Applications and Food Related Applications*. PhD thesis, Universitat de Barcelona, <http://hdl.handle.net/2445/69277>.
- Halang, W. A., Langlais, R., and Kugler, E. (1978). Cubic spline interpolation for the calculation of retention indices in temperature-programmed gas-liquid chromatography. *Analytical Chemistry*, 50(13):1829–1832, ISSN: 0003-2700, DOI: 10.1021/ac50035a026, <http://pubs.acs.org/doi/abs/10.1021/ac50035a026>.
- Harrington, P. d. B., Reese, E. S., Rauch, P. J., Hu, L., and Davis, D. M. (1997). Interactive Self-Modeling Mixture Analysis of Ion Mobility Spectra. *Appl. Spectrosc.*, 51(6):808–816, <http://as.osa.org/abstract.cfm?URI=as-51-6-808>.
- Hauschild, A.-C., Schneider, T., Pauling, J., Rupp, K., Jang, M., Baumbach, J. I., and Baumbach, J. (2012). Computational methods for metabolomic data analysis of ion mobility spectrometry data—reviewing the state of the art. *Metabolites*, 2(4):733–55, ISSN: 2218-1989, DOI: 10.3390/metabo2040733, <http://www.mdpi.com/2218-1989/2/4/733/htm>.
- Hoffmann, N. and Stoye, J. (2012). Generic Software Frameworks for GC-MS Based Metabolomics. *Metabolomics*, 2(2):73–98, ISSN: 1573-3882, 1573-3890, DOI: 10.1007/s11306-006-0022-6, [http://cdn.intechopen.com/pdfs/28004/InTech-Generic\\_software\\_frameworks\\_for\\_gc\\_ms\\_based\\_metabolomics.pdf](http://cdn.intechopen.com/pdfs/28004/InTech-Generic_software_frameworks_for_gc_ms_based_metabolomics.pdf).

- Hyman, J. M. (1983). Accurate Monotonicity Preserving Cubic Interpolation. *SIAM Journal on Scientific and Statistical Computing*, 4(4):645–654, ISSN: 0196-5204, DOI: 10.1137/0904045, <http://epubs.siam.org/doi/10.1137/0904045>.
- Hyvärinen, A., Karhune, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, ISBN: 0471464198, <http://books.google.com/books?hl=en&lr=&id=96D0ypDwAkkC&pgis=1>.
- Ioannidis, J. P. a. and Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science*, 334(6060):1230–1232, ISBN: 0036-8075, ISSN: 1095-9203, DOI: 10.1126/science.1211811, <http://www.ncbi.nlm.nih.gov/pubmed/22144616>.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-Image Translation with Conditional Adversarial Networks. <http://arxiv.org/abs/1611.07004>.
- Karasek, F. W. and Denney, D. W. (1974). Detection of 2,4,6-trinitrotoluene vapours in air by plasma chromatography. *Journal of Chromatography A*, 93(1):141–147, ISBN: 0021-9673, ISSN: 00219673, DOI: 10.1016/S0021-9673(00)83025-3.
- Karpas, Z. (2000). Ion Mobility Spectrometry in Forensic Science. In *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, Chichester, UK, DOI: 10.1002/9780470027318.a1113, <http://doi.wiley.com/10.1002/9780470027318.a1113>.
- Karpas, Z. (2013). Applications of ion mobility spectrometry (IMS) in the field of foodomics. *Food Research International*, 54(1):1146–1151, ISSN: 09639969, DOI: 10.1016/j.foodres.2012.11.029, <http://linkinghub.elsevier.com/retrieve/pii/S096399691200498X>.
- Karpas, Z., Cohen, G., Atweh, E., Barnard, G., and Golan, M. (2012a). Recent Applications of Ion Mobility Spectrometry in Diagnosis of Vaginal Infections. *International Journal of Spectroscopy*, 2012:1–6, ISSN: 1687-9449, DOI: 10.1155/2012/323859, <http://www.hindawi.com/journals/ij/s/2012/323859/>.
- Karpas, Z., Guamán, A. V., Calvo, D., Pardo, A., and Marco, S. (2012b). The potential of ion mobility spectrometry (IMS) for detection of 2,4,6-trichloroanisole (2,4,6-TCA) in wine. *Talanta*, 93:200–5, ISSN: 1873-3573, DOI: 10.1016/j.talanta.2012.02.012, <http://www.sciencedirect.com/science/article/pii/S0039914012001208>.

- Karpas, Z., Pollevoy, Y., and Melloul, S. (1991). Determination of bromine in air by ion mobility spectrometry. *Analytica Chimica Acta*, 249(2):503–507, ISSN: 00032670, DOI: 10.1016/S0003-2670(00)83025-1, <http://www.sciencedirect.com/science/article/pii/S0003267000830251>.
- Kaye, W. J. and Stimac, R. M. (2015). Miniaturized ion mobility spectrometer.
- Khalid, T., Aggio, R., White, P., De Lacy Costello, B., Persad, R., Al-Kateb, H., Jones, P., Probert, C. S., and Ratcliffe, N. (2015). Urinary Volatile Organic Compounds for the Detection of Prostate Cancer. *PLOS ONE*, 10(11):e0143283, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0143283, <http://dx.plos.org/10.1371/journal.pone.0143283>.
- Kováts, E. (1958). Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, 41(7):1915–1932, ISSN: 0018019X, DOI: 10.1002/hlca.19580410703, <http://doi.wiley.com/10.1002/hlca.19580410703>.
- Laphorn, C., Pullen, F., and Chowdhry, B. Z. (2013). Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: Separating and assigning structures to ions. *Mass Spectrometry Reviews*, 32(1):43–71, ISSN: 02777037, DOI: 10.1002/mas.21349, <http://doi.wiley.com/10.1002/mas.21349>.
- Lawton, W. H. and Sylvestre, E. A. (1971). Self Modeling Curve Resolution. *Technometrics*, 13(3), DOI: 10.1080/00401706.1971.10488823, [http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1971.10488823#.U7a2g\\_4\\_kxA](http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1971.10488823#.U7a2g_4_kxA).
- Macías, S. (2017). *Tools for the pre-processing and analysis of GC/MS metabolomics data. Application to data from prostate cancer patients*. PhD thesis, Universitat de Barcelona.
- Maeder, M. (1987). Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry*, 59(3):527–530, ISSN: 0003-2700, DOI: 10.1021/ac00130a035, <http://dx.doi.org/10.1021/ac00130a035>.
- Márquez-Sillero, I., Cárdenas, S., Sielemann, S., and Valcárcel, M. (2014). On-line headspace-multicapillary column-ion mobility spectrometry hyphenation as a tool for the determination of off-flavours in foods. *Journal of Chromatography A*, 1333:99–105, ISBN: 0021-9673, ISSN: 00219673, DOI: 10.1016/j.chroma.2014.01.062.
- Márquez-Sillero, I., Cárdenas, S., and Valcárcel, M. (2011). Direct determination of 2,4,6-trichloroanisole in wines by single-drop ionic liquid microextraction coupled with multicapillary column separation and ion mobility spectrometry detection. *Journal of Chromatography A*, 1218(42):7574–7580, ISSN:

- 00219673, DOI: 10.1016/j.chroma.2011.06.032, <http://linkinghub.elsevier.com/retrieve/pii/S0021967311008442>.
- Marr, A. J. and Groves, D. M. (2003). Ion mobility spectrometry of peroxide explosives TATP and HMTD. *Int. J. Ion Mobil. Spectrom*, 6:59–62.
- McNair, H. M. and Miller, J. M. (2009). *Basic Gas Chromatography*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, ISBN: 9780470480106, DOI: 10.1002/9780470480106, <http://doi.wiley.com/10.1002/9780470480106>.
- Mehran, M., Cooper, W. J., Golkar, N., Nickelsen, M. G., Mittlefehldt, E. R., Guthrie, E., and Jennings, W. (1991). Elution order in gas chromatography. *Journal of High Resolution Chromatography*, 14(11):745–750, ISSN: 0935–6304, DOI: 10.1002/jhrc.1240141109, <http://doi.wiley.com/10.1002/jhrc.1240141109>.
- Mills, G. A. and Walker, V. (2001). Headspace solid-phase microextraction profiling of volatile compounds in urine: Application to metabolic investigations. *Journal of Chromatography B: Biomedical Sciences and Applications*, 753(2):259–268, ISSN: 13872273, DOI: 10.1016/S0378–4347(00)00554–5.
- Miron, S., Dossot, M., Carteret, C., Margueron, S., and Brie, D. (2011). Joint processing of the parallel and crossed polarized Raman spectra and uniqueness in blind nonnegative source separation. *Chemometrics and Intelligent Laboratory Systems*, 105(1):7–18, ISSN: 01697439, DOI: 10.1016/j.chemolab.2010.10.005, <http://www.sciencedirect.com/science/article/pii/S0169743910002017>.
- Montoliu, I., Tauler, R., Padilla, M., Pardo, A., and Marco, S. (2010). Multivariate curve resolution applied to temperature-modulated metal oxide gas sensors. *Sensors and Actuators B: Chemical*, 145(1):464–473, ISSN: 09254005, DOI: 10.1016/j.snb.2009.12.051, <http://www.sciencedirect.com/science/article/pii/S0925400509010041>.
- Mukhopadhyay, R. (2004). Don't Waste Your Breath. *Analytical Chemistry*, 76(15):273 A–276 A, DOI: 10.1021/ac041600+, <http://pubs.acs.org/doi/abs/10.1021/ac041600%2B>.
- Nanji, A. A., Lawrence, A. H., and Mikhael, N. Z. (1987). Use of Skin Surface Sampling and Ion Mobility Spectrometry as a Preliminary Screening Method for Drug Detection in an Emergency Room. *Journal of Toxicology: Clinical Toxicology*, 25(6):501–515, ISSN: 0731–3810, DOI: 10.3109/15563658708992653, <http://www.tandfonline.com/doi/full/10.3109/15563658708992653>.

- Newey, W. K. and Powell, J. L. (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55(4):819, ISSN: 00129682, DOI: 10.2307/1911031, <http://www.jstor.org/stable/1911031?origin=crossref>.
- Nič, M., Jirát, J., Košata, B., Jenkins, A., and McNaught, A., editors (2009). *IUPAC Compendium of Chemical Terminology*. IUPAC, Research Triangle Park, NC, ISBN: 0-9678550-9-8, DOI: 10.1351/goldbook, <http://goldbook.iupac.org>.
- Nielsen, N. N.-P. V., Carstensen, J. M. J., and Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17-35, ISSN: 0021-9673, DOI: 10.1016/S0021-9673(98)00021-1, <http://www.sciencedirect.com/science/article/pii/S0021967398000211>[http://pdn.sciencedirect.com/science?\\_ob=MiamiImageURL&\\_cid=271409&\\_user=145085&\\_pii=S0021967398000211&\\_check=y&\\_origin=article&\\_zone=toolbar&\\_coverDate=1998-01&view=c&originContentFamily](http://pdn.sciencedirect.com/science?_ob=MiamiImageURL&_cid=271409&_user=145085&_pii=S0021967398000211&_check=y&_origin=article&_zone=toolbar&_coverDate=1998-01&view=c&originContentFamily).
- Nuzillard, D. and Nuzillard, J.-M. (1998). Application of blind source separation to 1-D and 2-D nuclear magnetic resonance spectroscopy. *IEEE Signal Processing Letters*, 5(8):209-211, ISSN: 1070-9908, DOI: 10.1109/97.704974, <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=704974>.
- O'Callaghan, S., De Souza, D. P., Isaac, A., Wang, Q., Hodkinson, L., Olshansky, M., Erwin, T., Appelbe, B., Tull, D. L., Roessner, U., Bacic, A., McConville, M. J., Likić, V. A., DeSouza, D. P., Isaac, A., Wang, Q., Hodkinson, L., Olshansky, M., Erwin, T., Appelbe, B., Tull, D. L., Roessner, U., Bacic, A., McConville, M. J., and Likic, V. A. (2012). PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC bioinformatics*, 13(1):115, ISSN: 1471-2105, DOI: 10.1186/1471-2105-13-115, <http://www.biomedcentral.com/1471-2105/13/115>, <http://www.biomedcentral.com/1471-2105/13/115/abstract><http://www.biomedcentral.com/1471-2105/13/115>.
- Oller-Moreno, S., Pardo, A., Jiménez-Soto, J. M., Samitier, J., and Marco, S. (2014). Adaptive Asymmetric Least Squares baseline estimation for analytical instruments. In Kanoun, O., editor, *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, pages 1-5, Castelldefels-Barcelona, Spain. IEEE, ISBN: 978-1-4799-3866-7, DOI: 10.1109/SSD.2014.6808837, <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6808837>.
- O'Donnell, R. M., Sun, X., and Harrington, P. d. B. (2008). Pharmaceutical applications of ion mobility spectrometry. *TrAC Trends in Analytical Chem-*



- istry*, 27(1):44–53, ISSN: 01659936, DOI: 10.1016/j.trac.2007.10.014, <http://linkinghub.elsevier.com/retrieve/pii/S0165993607002427>.
- Pasti, L., Walczak, B., Massart, D. L., and Reschiglian, P. (1999). Optimization of signal denoising in discrete wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, 48(1):21–34, ISBN: 0169-7439, ISSN: 01697439, DOI: 10.1016/S0169-7439(99)00002-7, <http://linkinghub.elsevier.com/retrieve/pii/S0169743999000027>.
- Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., and Tan, J. (2010). Asymmetric least squares for multiple spectra baseline correction. *Analytica Chimica Acta*, 683(1):63–68, ISSN: 00032670, DOI: 10.1016/j.aca.2010.08.033, <http://linkinghub.elsevier.com/retrieve/pii/S0003267010010627>.
- Phillips, M., Basa-Dalay, V., Bothamley, G., Cataneo, R. N., Lam, P. K., Natividad, M. P. R., Schmitt, P., and Wai, J. (2010). Breath biomarkers of active pulmonary tuberculosis. *Tuberculosis*, 90(2):145–151, ISSN: 14729792, DOI: 10.1016/j.tube.2010.01.003, <http://linkinghub.elsevier.com/retrieve/pii/S1472979210000156>.
- Phillips, M., Cataneo, R. N., Ditkoff, B. A., Fisher, P., Greenberg, J., Gunawardena, R., Kwon, C. S., Tietje, O., and Wong, C. (2006). Prediction of breast cancer using volatile biomarkers in the breath. *Breast Cancer Research and Treatment*, 99(1):19–21, ISSN: 0167-6806, DOI: 10.1007/s10549-006-9176-1, <http://link.springer.com/10.1007/s10549-006-9176-1>.
- 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, ISSN: 1471-2105, DOI: 10.1186/1471-2105-11-395, <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-395>.
- Pomareda, V., Calvo, D., Pardo, A., and Marco, S. (2010). Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra. *Chemometrics and Intelligent Laboratory Systems*, 104(2):318–332, ISSN: 01697439, DOI: 10.1016/j.chemolab.2010.09.010, <http://www.sciencedirect.com/science/article/pii/S0169743910001772>.
- Puton, J., Nousiainen, M., and Sillanpää, M. (2008). Ion mobility spectrometers with doped gases. *Talanta*, 76(5):978–987, ISSN: 00399140, DOI: 10.1016/j.talanta.2008.05.031, <http://linkinghub.elsevier.com/retrieve/pii/S003991400800430X>.

- Raatikainen, O., Reinikainen, V., Minkkinen, P., Ritvanen, T., Muje, P., Purssiainen, J., Hiltunen, T., Hyvönen, P., Von Wright, A., and Reinikainen, S. P. (2005). Multivariate modelling of fish freshness index based on ion mobility spectrometry measurements. In *Analytica Chimica Acta*, volume 544, pages 128–134. ISBN: 0003-2670, ISSN: 00032670, DOI: 10.1016/j.aca.2005.02.029.
- Räsänen, R.-M., Nousiainen, M., Peräkörpi, K., Sillanpää, M., Polari, L., Anttalainen, O., and Utriainen, M. (2008). Determination of gas phase triacetone triperoxide with aspiration ion mobility spectrometry and gas chromatography–mass spectrometry. *Analytica Chimica Acta*, 623(1):59–65, ISSN: 00032670, DOI: 10.1016/j.aca.2008.05.076, <http://linkinghub.elsevier.com/retrieve/pii/S0003267008010490>.
- Rendeiro, A. F., Schmidl, C., Strefford, J. C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., and Bock, C. (2016). Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nature Communications*, 7:11938, ISSN: 2041-1723, DOI: 10.1038/ncomms11938, <http://www.nature.com/doifinder/10.1038/ncomms11938>.
- Robinson, M. D., De Souza, D. P., Keen, W. W., Saunders, E. C., McConville, M. J., Speed, T. P., and Likić, V. a. (2007). A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC bioinformatics*, 8:419, ISSN: 1471-2105, DOI: 10.1186/1471-2105-8-419, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2194738&tool=pmcentrez&rendertype=abstract>.
- Salit, M. L. and Turk, G. C. (1998). A drift correction procedure. *Analytical chemistry*, 70(15):3184–3190.
- Sarker, S. D. and Nahar, L. (2012). Hyphenated Techniques and Their Applications in Natural Products Analysis. In Sarker, S. D. and Nahar, L., editors, *Natural Products Isolation*, pages 301–340. Humana Press, Totowa, NJ, ISBN: 978-1-61779-624-1, DOI: 10.1007/978-1-61779-624-1\_12, [http://dx.doi.org/10.1007/978-1-61779-624-1\\_12](http://dx.doi.org/10.1007/978-1-61779-624-1_12).
- Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, ISSN: 0003-2700, DOI: 10.1021/ac60214a047, <http://dx.doi.org/10.1021/ac60214a047>.
- Savorani, F., Tomasi, G., and Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 202(2):190–202, ISSN: 1096-0856, DOI:

- 10.1016/j.jmr.2009.11.012, <http://www.sciencedirect.com/science/article/pii/S1090780709003334>.
- Shao, X.-G., Leung, A. K.-M., and Chau, F.-T. (2003). Wavelet: A New Trend in Chemistry. *Accounts of Chemical Research*, 36(4):276–283, ISSN: 0001-4842, DOI: 10.1021/ar990163w, <http://pubs.acs.org/doi/abs/10.1021/ar990163w>.
- Shellie, R. A., Welthagen, W., Zrostliková, J., Spranger, J., Ristow, M., Fiehn, O., and Zimmermann, R. (2005). Statistical methods for comparing comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts. *Journal of Chromatography A*, 1086(1-2):83–90, ISSN: 00219673, DOI: 10.1016/j.chroma.2005.05.088, <http://linkinghub.elsevier.com/retrieve/pii/S0021967305011593>.
- Smit, S., van Breemen, M. J., Hoefsloot, H. C., Smilde, A. K., Aerts, J. M., and de Koster, C. G. (2007). Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta*, 592(2):210–217, ISSN: 00032670, DOI: 10.1016/j.aca.2007.04.043, <http://linkinghub.elsevier.com/retrieve/pii/S0003267007007726>.
- 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, ISSN: 0003-2700, DOI: 10.1021/ac051437y, <http://dx.doi.org/10.1021/ac051437y>.
- Sysi-Aho, M., Katajamaa, M., Yetukuri, L., and Oresic, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics*, 8(1):93, ISBN: 1471-2105 (Electronic), ISSN: 1471-2105, DOI: 10.1186/1471-2105-8-93, <http://www.biomedcentral.com/1471-2105/8/93>.
- Szymańska, E., Brodrick, E., Williams, M., Davies, A. N., Van Manen, H. J., and Buydens, L. M. C. (2015). Data size reduction strategy for the classification of breath and air samples using multicapillary column-ion mobility spectrometry. *Analytical Chemistry*, 87(2):869–875, ISSN: 15206882, DOI: 10.1021/ac503857y.
- Tauler, R. (1995). Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems*, 30(1):133–146, ISSN: 01697439, DOI: 10.1016/0169-7439(95)00047-X, <http://www.sciencedirect.com/science/article/pii/016974399500047X>.

- Thompson, I. M., Pauler, D. K., Goodman, P. J., Tangen, C. M., Lucia, M. S., Parnes, H. L., Minasian, L. M., Ford, L. G., Lippman, S. M., Crawford, E. D., Crowley, J. J., and Coltman, C. a. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level lower or equal to 4.0 ng per milliliter. *The New England journal of medicine*, 350(22):2239–2246, ISSN: 1533-4406, DOI: 10.1056/NEJMoa031918, <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&amp;id=15163773&retmode=ref&cmd=prlinks>.
- 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, ISSN: 00219673, DOI: 10.1016/j.chroma.2011.08.086, <http://www.sciencedirect.com/science/article/pii/S0021967311013148>  
[http://pdn.sciencedirect.com/science?\\_ob=MiamiImageURL&\\_cid=271409&\\_user=145085&\\_pii=S0021967311013148&\\_check=y&\\_origin=article&\\_zone=toolbar&\\_coverDate=2011--28&view=c&originContentFamily](http://pdn.sciencedirect.com/science?_ob=MiamiImageURL&_cid=271409&_user=145085&_pii=S0021967311013148&_check=y&_origin=article&_zone=toolbar&_coverDate=2011--28&view=c&originContentFamily).
- 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, ISSN: 0886-9383, DOI: 10.1002/cem.859, <http://doi.wiley.com/10.1002/cem.859>.
- Vaidya, R. A. and Hester, R. D. (1984). Deconvolution of overlapping chromatographic peaks using constrained non-linear optimization. *Journal of Chromatography A*, 287:231–244, ISSN: 00219673, DOI: 10.1016/S0021-9673(01)87700-1, <http://linkinghub.elsevier.com/retrieve/pii/S0021967301877001>.
- Vautz, W., Baumbach, J. I., and Jung, J. (2006a). Beer Fermentation Control Using Ion Mobility Spectrometry - Results of a Pilot Study. *Journal of the Institute of Brewing*, 112(2):157–164, ISSN: 00469750, DOI: 10.1002/j.2050-0416.2006.tb00245.x, <http://doi.wiley.com/10.1002/j.2050-0416.2006.tb00245.x>.
- Vautz, W., Baumbach, J. I., Jung, J., and Brew, J. I. (2006b). Beer Fermentation Control Using Ion Mobility Spectrometry – Results of a Pilot Study. *Journal Of The Institute Of Brewing*, 112(2):157–164, ISSN: 00469750, DOI: 10.1002/j.2050-0416.2006.tb00245.x.
- Vautz, W., Zimmermann, D., Hartmann, M., Baumbach, J. I., Nolte, J., and Jung, J. (2006c). Ion mobility spectrometry for food quality and safety. *Food Additives and Contaminants*, 23(11):1064–1073, ISSN: 0265-203X, DOI: 10.1080/02652030600889590, <http://www.tandfonline.com/doi/abs/10.1080/02652030600889590>.

- Vu, T. and Laukens, K. (2013). Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites*, 3(2):259–276, ISSN: 2218-1989, DOI: 10.3390/metabo3020259, <http://www.mdpi.com/2218-1989/3/2/259/>.
- Wentzell, P. D. and Brown, C. D. (2000). Signal Processing in Analytical Chemistry. In *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, Chichester, UK, DOI: 10.1002/9780470027318.a5207, <http://doi.wiley.com/10.1002/9780470027318.a5207>.
- Westhoff, M., Litterst, P., Maddula, S., Bödeker, B., Rahmann, S., Davies, A. N., and Baumbach, J. I. (2010). Differentiation of chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control group by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, 13(3):131–139, ISSN: 14356163, DOI: 10.1007/s12127-010-0049-2.
- Windig, W., Gallagher, N. B., Shaver, J. M., and Wise, B. M. (2005). A new approach for interactive self-modeling mixture analysis. *Chemometrics and Intelligent Laboratory Systems*, 77(1-2):85–96, ISSN: 01697439, DOI: 10.1016/j.chemolab.2004.06.009, <http://www.sciencedirect.com/science/article/pii/S0169743904002011>.
- Windig, W. and Guilment, J. (1991). Interactive self-modeling mixture analysis. *Analytical Chemistry*, 63(14):1425–1432, ISSN: 0003-2700, DOI: 10.1021/ac00014a016, <http://dx.doi.org/10.1021/ac00014a016>.
- Worley, B. and Powers, R. (2012). Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1):92–107, ISBN: 1863-0650, ISSN: 2213235X, DOI: 10.2174/2213235X11301010092, <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=2213-235X&volume=1&issue=1&spage=92>.
- Xia, J., Broadhurst, D. I., Wilson, M., and Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics : Official journal of the Metabolomic Society*, 9(2):280–299, ISSN: 1573-3882, DOI: 10.1007/s11306-012-0482-9, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3608878&tool=pmcentrez&rendertype=abstract>.
- Zamora, D., Alcalà, M., and Blanco, M. (2011). Determination of trace impurities in cosmetic intermediates by ion mobility spectrometry. *Analytica Chimica Acta*, 708(1-2):69–74, ISSN: 00032670, DOI: 10.1016/j.aca.2011.09.035, <http://linkinghub.elsevier.com/retrieve/pii/S0003267011013018>.

- Zhang, Z.-M., Chen, S., and Liang, Y.-Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *The Analyst*, 135(5):1138, ISSN: 0003-2654, DOI: 10.1039/b922045c, <http://xlink.rsc.org/?DOI=b922045c>.