Signal Processing Approaches to the Detection and Localization of Gas Chemical Sources using Partially Selective Sensors

Víctor Pomareda Sesé

FACULTAT DE FÍSICA
Departament d'Electrònica

MEMÒRIA PER OPTAR AL TÍTOL DE DOCTOR PER LA UNIVERSITAT DE BARCELONA

Doctorat en Enginyeria i Tecnologies Avançades (RD 99/2011)

_____

Signal Processing Approaches to the Detection and Localization of Gas Chemical Sources using Partially Selective Sensors

by

Víctor Pomareda Sesé

_____

Director i tutor:
Dr. Santiago Marco Colás

Barcelona, 5 de juny de 2013

Agraïments

Primer de tot voldria agrair als meus pares, germans, familiars i amics per tot el seu suport incondicional, per haver cregut sempre en mi. Sense ells no seria jo. Moltes gràcies, us estimo.

Després voldria agrair al Dr. Santiago Marco per ser el director de la meva tesis i haver-me guiat i aconsellat durant aquest llarg procés que quasi ha durat cinc anys, per haver pensat en mi en múltiples ocasions per assistir a cursos i conferències internacionals. Moltes gràcies.

Per suposat, també agraeixo a tots aquells que han estat els meus companys de feina i que s'han comportat com a tal: Toni, Agustín, Dani, Idoya, Aina, Miriam, Jordi, Fran, Eduard, Marta, Benja, Erola, Didier, Lucas, Davide, Masouhme, Bernd, Selda, Lechón, Antonio, Pablo, Sergis, Rafa (Dani), Juanma, Nil, Ariadna, Milad... i espero no deixar-me ningú...
D'entre aquests companys, voldria donar un agraïment especial al Luis Fernández i a la Ana Guamán, amb els quals vam començar la tesis a la mateixa època i portem tots aquests anys compartint experiències. Els menciono especialment perquè crec que s'han portat amb mi com alguna cosa més que companys, donant-me sempre suport en tot moment i animant-me en els moments difícils. Moltes gràcies.

Una experiència interessant que vaig poder gaudir durant la tesis va ser la realització d'una estada de 4 mesos a Örebro (Suècia). Allà vaig conéixer a gent molt interessant i vaig fer alguns amics que em van ajudar a integrar-me i a sentir-me a gust dins del grup. Vull agrair especialment al Dr. Achim Lilienthal per haver-me acollit, al Dr. Marco Trincavelli per haver-me supervisat, i al Víctor Hernández i al Ali Abdul Khaliq per haver estat els meus companys de feina durant l'estada i sense els quals els experiments amb el gran RASMUS (gaudint ara d'una merescuda jubilació) no haguessin estat possibles. Moltes gràcies.

També voldria agrair a les diferents administracions i institucions públiques i privades (tot i la retallada de la paga de desembre 2012...) pel seu suport econòmic sense el qual la realització de la tesis hagués estat impossible: RAMEM, *European Comission*, AGAUR, FBG, IBEC.

I per últim, finalment, voldria agrair molt especialment a tots aquells que faran un esforç per llegir-se la tesis i no només els agraïments i/o el resum. Gràcies per valorar l'esforç i el treball dedicat.

Abstract

Due to recent progress, higher-order chemical instrumentation provides large amounts of data which need automated processing in order to extract relevant information. In most cases, the raw signals or spectra are too complex for manual analysis. The ability to detect, identify and quantitate chemical substances in gas phase in field operations is required in a huge number of applications. Among them, I would like to highlight the need for chemical sensing on diverse humanitarian, safety and security applications. In these cases, it becomes extremely important to continuously monitor the environments where chemicals are spread in order to be ready to act when abnormal events are discovered.

In most critical scenarios the sample can not just be taken to the laboratory and analyzed, since an immediate answer is needed. In some other scenarios, the exploration of the area must be performed because the localization of the gas source or material of interest is unknown. This exploration can be performed using multiple mobile sensors in order to localize the chemical source or material.

Different sensing technologies have been successfully used to detect and identify different chemical substances (gases or volatile compounds). These compounds could be either toxic or hazardous, or they can be signatures of the materials to be detected, for instance, explosives or drugs.

Among these technologies, mobility based analyzers provide fast responses with high sensitivity. However, IMS instruments are not exempt of problems. Typically, they provide moderate selectivity, and overlapped peaks in the spectra are usual. Moreover, the presence of humidity makes peaks wider, thus worsening the resolving power and the resolution. Furthermore, the response of IMS is non-linear as substance concentration increases and more than one peak can appear in the spectra due to the same compound. Some authors advocate the use of multivariate data processing techniques for IMS, but further investigations are needed to set up the advantages and limits of this approach, compared to univariate alternatives.

In the present thesis, these problems are addressed and applications using an Ion Mobility Spectrometer (IMS) and a Differential Mobility Analyzer (DMA) are shown.

For the first time, multivariate data analysis tools have been applied to DMA. It is shown that DMA can be considered as a good instrumental approach for the detection of explosives, and the detection and quantitation of environmental sensitive VOCs, provided that the instrument spectra are properly pre-processed and subsequently analyzed by multivariate data processing.

Furthermore, Multivariate curve resolution Alternating Least Squares (MCR-ALS) is shown to be suitable to analyze IMS spectra qualitatively when interfering chemicals appear in the spectra and even when their behaviour is non-linear. Partial Least Squares (PLS) methods are demonstrated to work properly for the quantitative analysis of these signals.

It is also demonstrated in this thesis that the quantitative measurements from these sensors can be integrated in a gas source localization algorithm in order to improve the localization of the source in those scenarios where it is required. It is shown that the new proposal works significantly better in cases where the source strength is weak. This is illustrated presenting results from simulations generated under realistic conditions.

Moreover, real-world data were obtained using a mobile robot mounting a photo ionization detector (PID). Experiments were carried out under forced ventilation and turbulences in indoors and outdoors environments. The results obtained validate the simulation results and confirm that the new localization algorithm can effectively operate in real environments.

Contents

Chapter 1

Chapter 2

Chapter 3

3 Multivariate techniques applied to a novel Differential Mobility Analyzer (DMA) for explosives detection and VOCs identification and quantitation

Chapter 4

4 Sparse MCR-ALS using L1 regularization and Gaussian peak shape models

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Chapter 1

Introduction

1.1 Motivation

As technology evolves, faster, smaller and more powerful processors are produced. Processors could take the form of microcontrollers, general purpose processors, digital signal processors, FPGAs or even embedded PCs. Processors can be integrated more easily in sensor systems and at lower cost; thus, when building instrumentation, incorporating intelligence becomes a standard practice.

Today instrumentation is digital and features at least firmware for sensor and actuator interface, data-logging, digital communications, and user interface. It is also possible to embed intelligence beyond the basic functionality mentioned before.

In analytical instrumentation, the ability to generate data is larger than the capacity to analyze it. Current instrumentation can produce large amounts of data, but the data are often complex, and manual analysis becomes either a bottleneck or it is just impossible. The analysis and interpretation of high throutput techniques requires automatic multivariate and data processing approaches to extract useful hidden information.

The ability to detect and correctly identify chemical substances in gas phase in field operations is required in a large number of applications; among them environmental monitoring, exploration of areas where hazardous chemicals have been spread, detection of explosives, localization of clandestine chemical laboratories, or humanitarian and security operations. Moreover, in these cases, it becomes extremely important to continuously monitor the environments where chemicals are spread in order to be ready to act when an abnormal event is produced.

In some scenarios the sample is taken manually and analyzed "in situ" or subsequently in a laboratory; but in many others, where the locations of the chemical sources are unknown, then area exploration is needed. If the chemical can be detected far from the source, but still the source location is unknown, we have today a number of algorithmic proposals to estimate the position of the source from the sequence of readings along the area being explored. In other occasions, the emission from the source is so weak (due to low vapour pressure of materials producing the vapours) that the detection has to be done in proximity. Exhaustive search is the only option, and to speed up explorations, extremely fast analysis is needed.
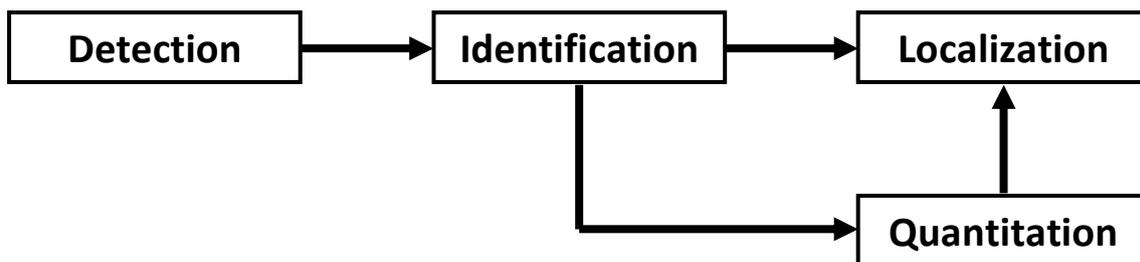
Figure 1.1. The ability to detect and identify chemical substances is required in many applications. Source localization is required if the source location is unknown. A substance quantitation step can help to solve the source localization problem.

Chemical substances, which can be all sorts of molecules from simple atoms to complex biological molecules, can be detected using different sensing technologies. Ideally one would like to have as many different kinds of sensors as the number of different analytes, being each kind of sensor perfectly selective to a specific analyte, but practically this is difficult to achieve and typically one chemical sensor reacts to different analytes instead. Even biosensors featuring good selectivity (e.g. antibodies) are known to have cross-sensitivities to structurally related compounds (Calvo et al. 2011). Moreover, it could be that a selective sensor for a certain substance does not exist, thus a non-selective sensor has to be used.

The main advantage of using non-selective sensors, which potentially is also a disadvantage, is that the sensors respond to a large number of different substances and this allows using the sensors in different applications, minimizing costs compared to using different sensors selective to specific substances, which are usually more expensive since in many occasions adhoc development is needed. In addition, in some cases a global odour assessment is required. This is typically the case when evaluating the organoleptic profile of food product. This profile could depend on tens, if not hundreds, of volatiles (Buffo & Cardelli-Freire 2004). Then it is not generally feasible to have as many selective sensors as volatiles present in the food headspace. However, when using non-selective sensors, since they react with many different substances, interfering responses can appear potentially masking the signal of interest, thus hindering the detection.

Non-selective sensors (as for instance Flame Ionization Detectors or Photo Ionization Detectors) can be combined with Chromatographic pre-separation. Correct optimization of the column (materials, temperatures, ramps, etc) can provide high selectivity at the expense of slower analysis (tens of minutes typically). In fact, it is well-known that speeding up the chromatographic pre-separation leads to a

degradation of the selectivity and makes co-elution more frequent. In most cases, chromatography is still too slow for the application scenarios described below.

When performing chemical source localization, and in order to detect the substance of interest far from the source, typically low thresholds are desired. However, low thresholds together with low selectivity lead to large number of false alarms that degrade most of the localization algorithms.

In order to deal with these issues, adequate signal processing techniques (adapted to the sensor technology) should be used to estimate qualitatively and quantitatively the sensor responses, so as to ameliorate the detection and localization of gas chemical sources.

Chemometrics can generally be described as the application of mathematical and statistical methods to improve chemical measurement processes, and to extract more useful chemical information from measurement data. Nowadays, it is common to approach problems realizing that there are deeply hidden relationships between variables that can be tackled only by the use of newer data analysis techniques; thus chemometrics becomes a necessity (Workman et al. 1996). With this regard, signal and data processing become essential elements in most chemical sensing instruments. The multivariate responses obtained by chemical sensor arrays and spectrometers require signal and data processing to carry out the fundamental tasks of odour identification (classification) and concentration estimation (regression). Since the paper published by Gutierrez-Osuna (Gutierrez-Osuna 2002), the basic methodology for building the data processing of the instruments is firmly established and important advances in the chemometrics field have shown that proper processing of the sensor signals can improve the robustness of the instruments against diverse perturbations, namely, environmental variables, background changes or drift (Marco & Gutierrez-Galvez 2012).

1.2 Scenarios

Scenarios where chemicals in vapour phase have to be detected and quantified are varied. Eventually, when the source location is unknown, it has to be localized. Moreover, the presence of interfering substances hinders carrying out these tasks. In this section, some scenarios which have important implications for security and safety issues are described.

1.2.1 Airport screening

In the aftermath of the tragic events of September 11, 2001 in the USA, numerous changes have been made to aviation security policy and operations throughout the world, especially in the USA.

In the past, the use of trained dogs was the preferred solution to the detection of explosives and drugs. Sniffer dogs have good mobility capabilities, they are fast, they are able to detect up to 9-14 explosives per dog and they achieve very low limits of detection; however the time of operation is usually limited (1 hour between breaks) and they need caring and special training (Furton & Myers 2001).

Current challenges in security applications and the last changes in aviation security policy incentivize an increasing demand in alternative solutions.

Figure 1.2. Sniffer dogs can detect a high number of different substances achieving low limits of detection. However, they need caring and special training. Image sources: (www.smh.com.au) and (www.bbc.co.uk).

Explosive detection systems (EDS) (Singh & Singh 2003) and explosive trace detection (ETD) (Moore 2007) are considered nowadays an essential component of aviation security strategies designed to prevent and deter terrorist threats. The allocation and utilization of checked baggage screening devices is a critical component in aviation security systems (Jacobson et al. 2005).

There is therefore an increasing focus on new technologies that can be applied to security screening, either to simplify or speed up the checking process, or to provide additional functionality. Reliable instrumentation able to operate with volatile samples, at trace levels, and with minor or inexistent sample preparation is required. High sensitivities and selectivities are also required to prevent the occurrence of false negatives and false positives.

Different technologies have been successfully applied to the detection of drugs and explosives, for instance: Mass Spectrometry (MS) in the detection of explosives analyzing solid surfaces (Na et al. 2007); Raman Spectrometry in the analysis of drugs (Hargreaves et al. 2008) and the detection of explosives and precursors on clothing (Ali, Edwards & Scowen 2009); Thermal-Neutron Analysis (TNA) for explosive detection in airline baggage (Shea, Gozani & Bozorgmanesh 1990); Terahertz Spectroscopy for explosives detection (Leahy-Hoppa, Fitch & Osiander 2009); Ultra-Low-Field Magnetic Resonance Imaging (ULF-MRI) to inspect liquids non-invasively at a security checkpoint for detecting hazardous material (Espy et al. 2010); Ion Mobility Spectrometry (IMS) (Eiceman & Karpas 2005) for drugs detection (Lai, Corbin & Almirall 2008) in human hair (Keller et al. 1998) or on the hands of subjects (Lawrence 1987), for explosive detection (Ewing et al. 2001; Ewing & Miller 2001; Lai et al. 2008; Hilton et al. 2010) and even for detecting TNT in the presence of interferents (Matz, Tornatore & Hill 2001); among many other technologies (Wallin et al. 2009; Singh & Singh 2003).

Ion Mobility Spectrometry (IMS) is probably one of the most used sensing technologies in airport security. Different commercial instruments in different configurations are available for this scenario, specially optimized to the detection of explosives and narcotics. For instance, Smiths Detection (Watford, UK) produced the *IonScan* which permits the simultaneous detection of different explosives (picogram range) and drugs (sub-nanogram range) in less than 8 seconds (IONSCAN) taking the sample from solid surfaces with a swap and then thermodesorbing the collected particles.

Since 2001, the possibility to develop a portal for passenger screening has attracted many companies. However, for exhaustive screening, analysis time has to be less than 1 minute putting enormous pressure on the sampling. With this regard, Smiths Detection manufactures the *Sentinel II*, which extracts and analyze particles from the hair, body, clothes and shoes of the subject generating an airflow (SentinelII). A similar instrument (EntryScan4) is commercialized by Morpho Detection (Paris, France). These solutions are mostly based on Ion Mobility Spectrometry due to its fast analysis times and good limits of detection for illicit substances.

Figure 1.3. Different IMS configurations for aviation security provided by Smiths Detection.

Figure 1.4. IMS portals for drug and narcotic detection provided by Morpho Detection.

In this thesis, different experiments were performed using an airport security checkpoint prototype for baggage inspection provided by RAMEM S.A (Madrid, Spain) (RAMEM) in order to study if different substances could be detected even in the presence of interfering chemicals. The details are given in chapter 4.

1.2.2 Improvised explosive detection

An Improvised Explosive Device (IED) is defined as a device placed or fabricated in an improvised manner incorporating destructive, lethal, noxious, pyrotechnic, or incendiary chemicals and designed to destroy, incapacitate, harass, or distract (O'Hara 2008).

An IED may be prepared with very limited technical means and limited expertise using chemicals available in the drugstore. These homemade explosives may be used by terrorist in crowded infrastructures such as public streets, the underground, sports arenas, shopping centres or museums, for instance.

Different methods of detection are employed: analyzing the electromagnetic spectrum, using unmanned aerial vehicles (UAVs) mounting electro-optical sensors, or through chemical sensors. From the chemical point of view, hyperspectral sensing, Raman spectroscopy and sodium-ion detectors to identify nitrogen-based explosives are used (O'Hara 2008).

In this scenario we consider the specific case in which a suspicious bag, suitcase or package has already been identified and should be closely inspected. The investigation could be carried out using the sensor technologies mentioned in section 1.2.1.

1.2.3 Uncontrolled release of hazardous compounds

Although some compounds have a legitimate use mainly due to their industrial applications, an uncontrolled release at great scale could lead to dramatic consequences due to their toxicity (Hu & Raymond 2004). Some of them have been used in the past by terrorist or for chemical war (Berkowitz et al. 2002). For instance: phosgene is a precursor in the manufacture of plastics; ammonia is very valuable in agricultural industries for plant growth; sulphuric acid is used in fertilizers, insecticides or the production of detergent among other applications; or chloropicrin which is a fumigant but very irritant. These types of substances are known, in this context, as Toxic Industrial Compounds (TIC).

TICs are defined by the Occupational Safety & Health Administration (OSHA) in the USA as industrial chemicals that are manufactured, stored, transported, and used throughout the world. They can be in the gas, liquid, or solid state. They can be chemical hazards (e.g., carcinogens, reproductive hazards, corrosives, or agents that affect the lungs or blood) or physical hazards (e.g., flammable, combustible, explosive, or reactive) which, if obtained by terrorists or caused to be released, may have extremely serious effects on exposed individuals. Many toxic industrial chemicals are highly toxic and may rapidly affect exposed individuals. Toxic industrial chemicals (whether as a gas, aerosol, or liquid) enter the body through inhalation, through the skin, or through digestion. The time that it takes for a toxic industrial chemical to begin working is dependent mainly on the route that the agent enters the body. Generally poisoning occurs more quickly if a chemical enters through the lungs (because of the ability of the agent to rapidly diffuse throughout the body).

Unlike TICs, which need to be used in large quantities to be toxic, chemical warfare agents (CWAs) have a toxic effect at very small amounts. The chemical is either

injurious or lethal. These chemicals have been used in war in the past and different variants can be identified (Disaster_Information_Management_Research_Center): nerve agents, asphyxiant and blood agents, blister agents or pulmonary agents among others. For instance in 1995, sarin, a colorless and odorless liquid nerve agent, was released on the Tokyo underground in five coordinated attacks, killing thirteen people and severely injuring fifty. Each perpetrator carried approximately 900ml of sarin in plastic bags; a single drop can kill an adult.

Fortunately, these substances can be detected using different technologies (Sferopoulos 2009), some examples are: IMS, Flame photometry, Infra-red (IR) spectroscopy, Raman spectroscopy, surface acoustic wave (SAW) or photo ionization detection (PID) among others.

1.2.4 Leakage detection

In this scenario hazardous substances released at local scale or at certain specific locations and typically by accident and not on purpose are considered.

Typically, the budget is limited and cheap sensors are required. In these cases using an array of non-selective sensors can be very useful. For instance, a sensor array of different kinds of gas sensors (also known as electronic nose) is demonstrated in (Perera et al. 2006) to be useful for sensing air quality in high pressure lines where oil from an air compressor can contaminate the lines if the protection filters are damaged. The application allows an online analysis and continuous sensing of the correct functioning of the filters.

In other safety-related applications, the detection of certain substances is critical. For instance, carbon monoxide is an odourless, colourless and non-irritating gas which is known as the silent killer. It is usually the result of incomplete combustion due to domestic burners and boilers, car exhausts, log or carbon-based stoves and chimneys, or fire at its initial stages. Thus, accidental poisoning can be caused by inadequate ventilation or obstructed furnaces. A system to detect fire using an array of three gas sensors is presented in (Kohl, Kelleter & Petig 2001). Data processing is applied looking at the correlation between the sensors and the development of the signals over time. Furthermore, it is shown in (Raub et al. 2000) that the installation of a static network of selective carbon monoxide detectors can be a way to compensate for the lack of sensitivity to this gas.

1.2.5 Gas detection using mobile sensors

A static network of gas sensors only provides sparse measurements at certain locations. Moreover, it is not always possible to have an operator taken manually

measurements at predefined locations or it is not safe for the operator due to the dangerousness of the released gases (TICs or CWA). This spatial and temporal sparsity in the measurements complicates the detection of leaks and ultimately, it causes a waste of resources.

With this regard, mobile robotics can contribute to mitigate the spatial sparsity of the measurement by providing a versatile system that is able to adaptively collect measurement at different locations.

In (Hernández-Bennetts et al. 2012) it is demonstrated the ability to detect and create a distribution map of methane in indoors and outdoors. Methane is not toxic; however, it is extremely flammable and may form explosive mixtures with air; moreover, it is an asphyxiant and may displace oxygen in an enclosed space. It is the main component of natural gas, a relatively potent greenhouse gas and probably the most abundant organic compound on earth. At room temperature and standard pressure, methane is a colourless and odourless gas. In the cited work, the environment is sensed remotely using a RMLD (Remote Methane Leak Detector). This detector is very selective to methane and belongs to the family of TDLAS (Tunable Diode Laser Absorption Spectroscopy).

In another application (Kim et al. 2011), a rapid, flexible and remote technology as LIBS (Laser-induced Breakdown Spectroscopy) is used for detecting leakages of boric acid in a nuclear power plant.

In other example (Distante, Indiveri & Reina 2009), it is shown the ability to detect multiple odour sources and differentiate between sources very close to one another, using an array of tin oxide chemical sensors mounted on a mobile robot which follows a predefined trajectory in a hazardous industrial site. The sensors provide a real-time olfactory map of the environment.

In (Trincavelli, Coradeschi & Loutfi 2009) an array of five semiconductor gas sensors is mounted in a mobile robot for continuous monitoring and online classification. Three different substances are correctly identified in this work using the same sensor array.

1.2.6 Gas sensor networks in an urban-like environment

Diverse causes may require chemical surveillance within an urban environment. In some cases it is just pollution monitoring, but in this thesis we are mostly considering critical applications related to the fight against terror.

There have been previous evidences of the presence of clandestine labs of explosives and most commonly labs of drugs in large cities. The chemical processes carried out in those labs produce exhaust chemicals that are dispersed either in the atmosphere or in the public sewage. For instance, it is known that explosions in London's terrorist attack in 2005 were caused by homemade organic peroxide-based devices packed into

backpacks and that these explosives had been being prepared during the previous weeks in clandestine kitchens. Some neighbours also reported that they noted degradation of the plants surrounding the areas during those weeks.

In an urban environment, it is necessary to detect explosives and their precursors during the production stage rather than preventing terrorist attacks while they are already in motion, which is extremely difficult. As pointed before (section 1.2.2), improvised explosives can be home-produced with minimal technology (kitchen-like). In the case of drug production, it is critical to find the illegal drug laboratory before the distribution of drugs on the black market. Fortunately, the production stage forms a window of opportunity, usually of substantial time duration (several weeks of active production), where it is possible to detect and pinpoint clandestine chemical laboratories; thus police and security forces can intervene opportunely.
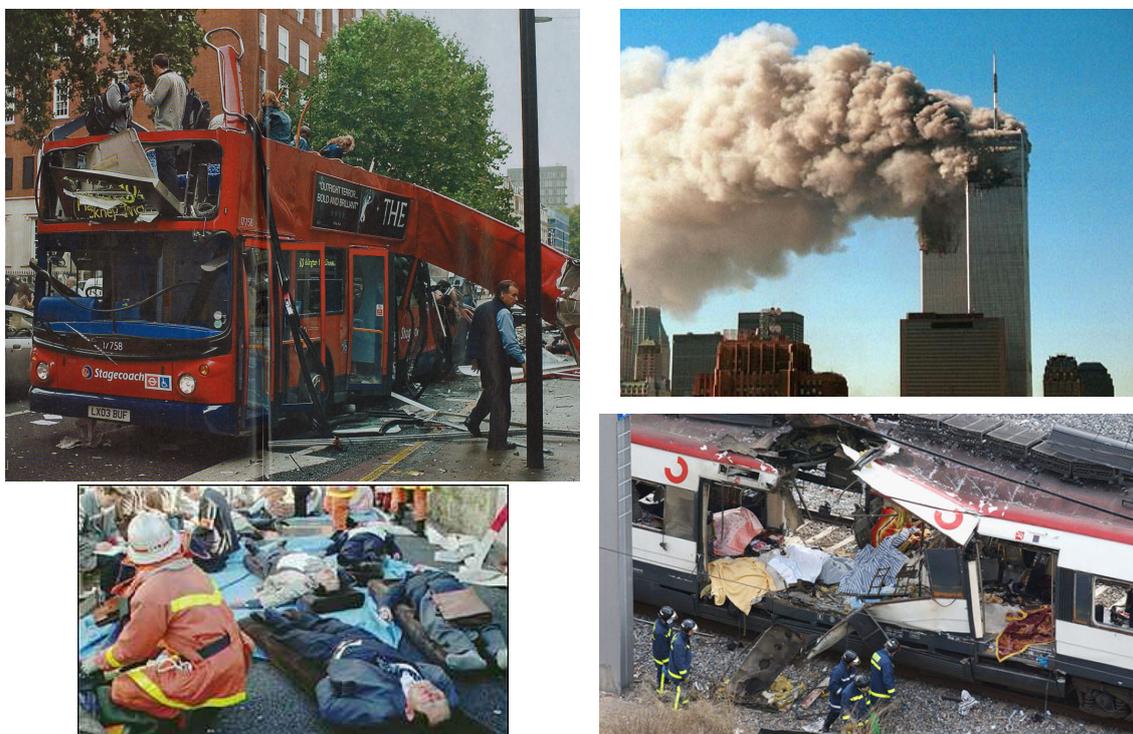
Figure 1.5. Terrorist attacks are extremely difficult to be stopped when they are already in motion. Explosives must be detected during the production stage. Image sources (from left to right and top to bottom): (www.julyseventh.co.uk), (www.bbc.co.uk), (news.bbc.co.uk) and (cnnespanol.cnn.com).

Because of the last big terrorist attacks in the last years (New York 2001, Madrid 2004 and London 2005), an increasing demand in reliable detection methods has appeared. Chemical precursors of drugs and explosives could be detected building a network of fixed and/or mobile sensors (typically included in police cars). These sensors continuously sample air while its carrier performs other tasks beyond chemical

surveillance. Sensor measurements together with GPS information can be transferred to a central system (using GPRS for instance) where all information is collected and stored. Then sensor data together with meteorological data (atmosphere stability, wind speed and direction, temperature, humidity...) can be processed and analyzed in order to identify and locate potential chemical sources.
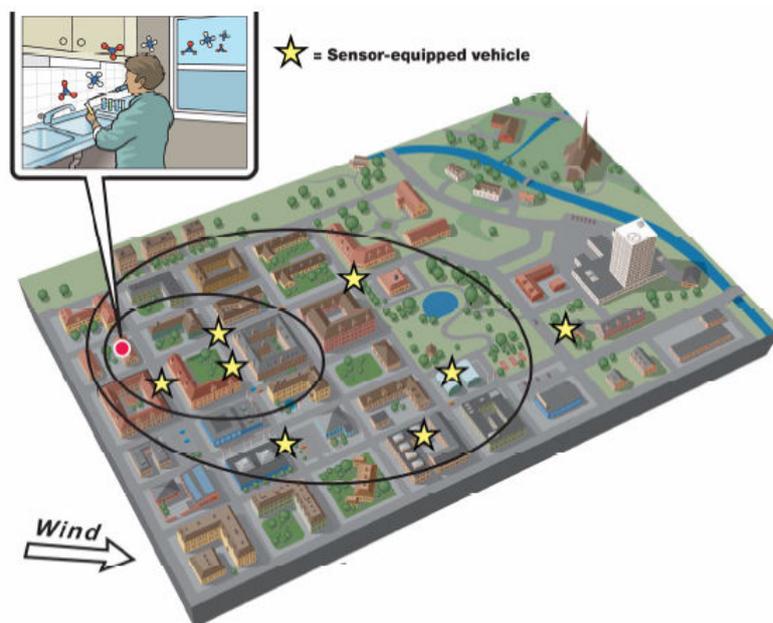
Figure 1.6. Scenario where there is a clandestine laboratory where illicit production of drugs or explosives takes place. A network of mobile sensors is deployed over the area. Real-time measurements by the sensors are sent to a central system where data is processed and analyzed.

Since there are hundreds of hazardous chemicals which could be potential threats and must be monitored, the use of partially selective sensors is encouraged. For instance, in (Abbaspour & Mansouri 2005) a network of fixed local sites is designed. Each local site contains a Combustible Gas Detector (CGD) and a Photo Ionization Detector (PID), which is a very sensitive and non-selective sensor that reacts with almost all chemicals in the air (section 1.3.4). Each local site communicates through a radio frequency modem to a local control panel forming a unit and each unit communicates to the central control panel where data from the sensors and triggered alarms are stored, processed and analyzed. In the central panel a map of the city can be visualized thus potential threats can be identified.

Another application when deploying gas sensor networks in this scenario is pollution monitoring in order to control air quality. Moreover, in some occasions, illegal emissions of pollutants also require source localization tasks. For instance, in (De Vito et al. 2011) and in (De Vito & Fattoruso 2012) preliminary results from a wireless

network of gas chemical sensors are presented. It is suggested that the system can effectively be used to detect and quantified pollutants in complex mixtures in real-time by applying trained artificial neural networks. Moreover, the three-dimensional distribution of gas is also estimated in real-time.

Since in these applications it is extremely important to localize the gas source, the distributed system has to face some problems which are still open issues. In an urban-like environment, dispersion of chemicals is more complex than in other open air scenarios, due to the complex topology and geometry of cities. The street structure leads in general to channelling effects and in consequence dispersion can change dramatically with small changes in wind direction. In addition to wind changes, meteorological conditions (atmosphere stability, temperature, humidity...) are also changing continuously and influence the dispersion. Finally, within a city we expect to find a background of interfering chemicals which evolves over space and time. All these problems are addressed in the present thesis and have a larger impact when the source strength is weak (this is the case for clandestine laboratories).

1.3 Gas sensing technologies

Different sensing technologies can be found in the literature and have been applied successfully to a huge number of different applications; however we focus here on the most relevant to the present thesis.

1.3.1 Ion Mobility Spectrometry

The term ion mobility spectrometry (IMS) (Hill et al. 1990) refers to the principles, methods and instrumentation for characterizing chemical substances on the basis of the velocity of gas-phase ions in an electric field (Eiceman & Karpas 2005).

IMS technology provides high-speed analysis (1-5 seconds), portability, no sample preparation (handheld devices), high sensitivity (ppb levels for some compounds) and comparatively low cost of operation. Moreover, in some instruments it is possible to invert the electric field polarity in order to analyze positive and negative ions separately (positive or negative mode). On the other hand, compared to Gas Chromatography (GC), IMS technology has only a moderate chemical selectivity and it gets worse when the instrument dimensions are scaled down (Babis et al. 2009), and also in the presence of humidity (Vautz, Sielemann & Baumbach 2004).

IMS technology was developed in the early 1970s and it has been mainly used in security applications such as chemical warfare agent detection (Kanu, Haigh & Hill 2005; G. A. Eiceman 2002; Rearden & Harrington 2005), screening for illicit substances (Lawrence 1987; Keller et al. 1998; Ochoa & Harrington 2004; Lai et al. 2008; Lai,

Corbin & Almirall 2008) and protection against explosives (Lai et al. 2008; Ewing et al. 2001; Buxton & Harrington 2003) and toxic compounds (Li et al. 2002; Utriainen, Kärpänoja & Paakkanen 2003). IMS has been also used in other scenarios such as environmental monitoring (Márquez-Sillero et al. 2011; Tuovinen, Paakkanen & Hänninen 2000; G.A Eiceman 2002) and pharmaceutical applications (O' Donnell, Sun & Harrington 2008; Strachan, Nicholson & Ogden 1995; Snyder, Blyth & Parsons 1996) and, gradually it is expanding its range of applications to food and beverage applications (Ogden & Strachan 1993; Vautz et al. 2006; Garrido-Delgado et al. 2011), clinical applications (Westhoff et al. 2007; Ruzsanyi & Baumbach 2005; Ruzsanyi et al. 2005; Westhoff et al. 2005) and industrial applications (G.A Eiceman 2002; Baumbach 2006; Eiceman et al. 1995; Lawrence, Barbour & Sutcliffe 1991), among many others (Borsdorf et al. 2011). With this regard, Armenta et al. published a review of recent an unconventional applications using IMS (Armenta, Alcala & Blanco 2011).

Desktop and also handheld instruments are available from a number of vendors; namely: Bruker Daltonics (Leipzig, Germany), Smiths Detection (Watford, England) (Figure 1.7), Airsense Analytics (Schwerin, Germany), Environics (Mikkeli, Finland), Thermo Fisher (Waltham, Massachusetts, USA). Examples of commercial instruments based on IMS are shown in Figure 1.3, Figure 1.4 and Figure 1.15(a).
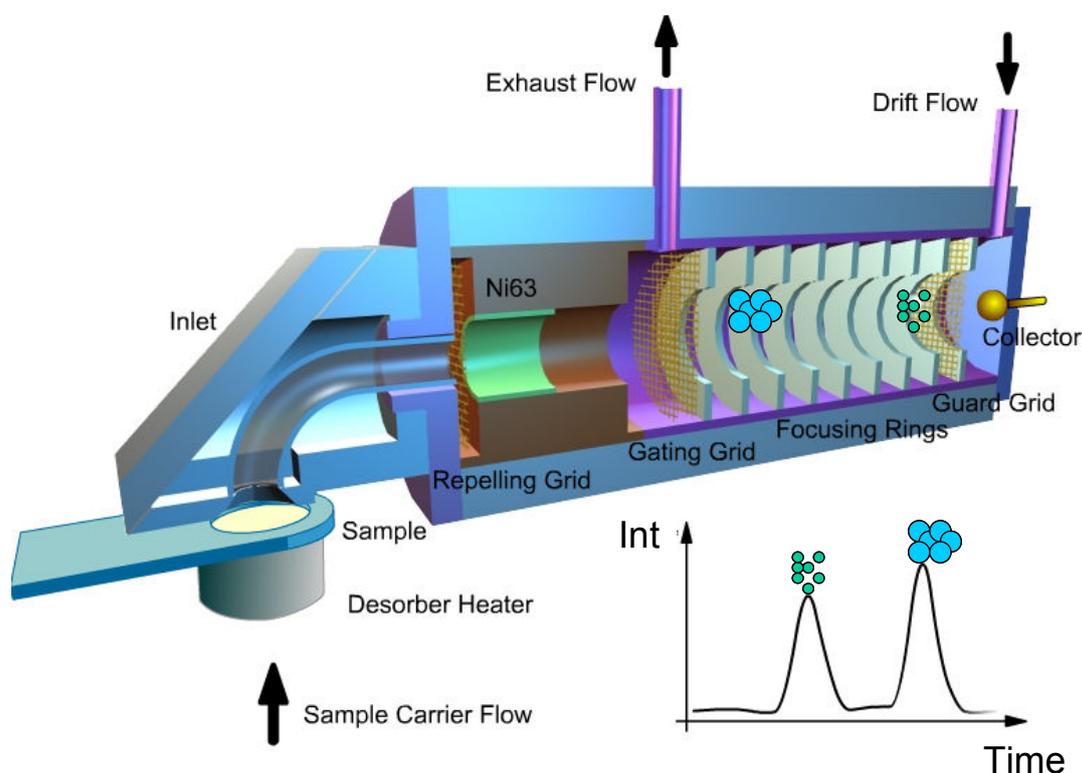
Figure 1.7. Schematic of a traditional ion mobility spectrometer with a linear electric field drift tube including a [63]Ni radioactive ionization source. Image adapted from (IONSCAN) (Smiths Detection).

The most important component of an IMS analyzer is the drift tube, since sensitivity and resolution depend on its design. There are two basic designs: the traditional configuration with a linear electric field drift tube (DC) and the field asymmetric design (AC or RF). Two main parts can be distinguished within the drift tube: the reactant region and the separation or drift region. In a traditional IMS with linear electric field, within the ionization chamber, gas phase molecules are ionized typically by a radioactive source, although a large variety of ionization sources are available, namely: UV-lamps (Baumbach et al. 2003), corona discharge (CD) (Sabo, Matúška & Matejčík 2011), electro-spray (ESI) (Wittmer et al. 1994; Hilton et al. 2010), or Matrix Assisted Laser Desorption Ionization (MALDI) (Chen 2008). Depending on the ionization source, volatile (radioactive, CD and UV), semi-volatile and non-volatile compounds (ESI and MALDI) can be analyzed. In Table 1.1 a comparison among ionization sources is given.

Source	Type of chemicals	Maintenance	Cost	Comments
Radioactive	Universal	Low	Medium/low	Licensing required
Corona discharge	Universal	High	Medium	Maintenance required
Photoionization	Selective	Medium	Medium	Low efficiency
Surface ionization	Nitrogen bases	High	Medium	Complex
Electrospray	Liquids	Medium	Medium	Long clearance
MALDI	Solids	High	High	Laboratory use
Flame	Selective	Medium	Low	Molecular structure lost

Table 1.1. Summary of ionization techniques used in ion mobility spectrometry. Extracted from (Eiceman & Karpas 2005).

When using ionization sources of high energy (radioactive, CD), the ionization chamber provides a reservoir of ions, known as reactant ions, for the incoming molecules. This reservoir produces the reactant ion peaks (RIP) in the spectra. When a sample is introduced into the spectrometer, chemical reactions at atmospheric pressure (APCI: Atmospheric Pressure Chemical Ionization (Bell et al. 1994)) take place and ion products are formed due to interactions between the reactant ions and the molecules in the sample. At this stage, the sensitivity of the instrument to the different analytes

in the sample is given by the affinity to form ionized clusters (proton affinity for positive ions and electronegativity for negative ions). The formed ion products in the reactant region are prevented from entering into the linear drift region by an electrostatic shutter grid acting as an ion gate (Shamlouei & Tabrizchi 2008). The more time the gate is closed, the more ion products accumulate and the better the sensitivity, but peaks get wider (less resolution) and time of analysis increases. When the gate grid opens, an electric field of typically 200V/cm accelerates ions into the drift region until they reach constant velocity due to collisions with the surrounding gas molecules. Inside the drift region there is a drift flow going on opposite direction which keep neutral species out to the tube. At the end of the drift region there is a collector which takes the charge from the ions neutralizing them and producing a current output. The collected charge is visualized as a mobility spectrum in only a few tens of milliseconds. This mobility (K) depends on the ion's size, shape and weight (Eiceman & Karpas 2005; Hill et al. 1990):

(Eq. 1.1)
$$K = \frac{3e(2\pi)^{1/2}(1+\alpha)}{16N\left(\mu k T_{eff}\right)^{1/2}\Omega_D(T_{eff})}$$

where e is the charge of an electron; N is the number density of neutral-gas molecules at the pressure of measurement; α is the correction factor; μ is the reduced mass of ion and gas of the supporting atmosphere; T_{eff} is the effective temperature of the ion determined by thermal energy and the energy acquired in the electric field, and Ω_D is the effective collision cross section of the ion in the supporting atmosphere.

Since K depends on the temperature and the supporting atmosphere (pressure); for identification purposes, the reduced mobility (K_0) is often used:

(Eq. 1.2)
$$K_0 = K \frac{P}{P_0} \frac{T_0}{T}$$

where typically, P_0 = 760mm Hg, and T_0 = 273K. The reduced mobility K_0 allows comparing results between different IMS instruments using the same ionization source.

In this traditional configuration, the mobility coefficient (K) is considered constant since the instrument operates in low electric fields (100-300V/cm). However, when the electric field is increased beyond the linear range (it depends on the analyte) it is observed that K becomes dependent on the electric field. This concept is exploited in the design of field asymmetric drift tubes. The technology is known as Differential

Mobility Spectrometry (DMS) or Field Asymmetric IMS (FAIMS) and different devices are based on this principle. For instance, in (Eiceman et al. 2004) a DMS is used for the separation of ions from explosives, or in (Krylov et al. 2002) the mobilities of ions related to different ketones are studied by Planar FAIMS (PFAIMS).

The ion separation capabilities of IMS are quantified using either the "resolving power" (R_p) or the "peak-to-peak resolution" (R_{pp}) formalisms (Spangler 2002), being these parameters defined respectively as:

(Eq. 1.3) $$R_p = t_d / w_h$$

where t_d is the drift time and w_h is the temporal full-width-at-half-height (FWHH) for the mobility peak;

(Eq. 1.4) $$R_{pp} = \frac{2(t_{d2} - t_{d1})}{1.7(w_{h2} - w_{h1})}$$

where subindices 1 and 2 refer to two neighbouring peaks.
Typical values for the resolving power (Eq. 1.3) in IMS instruments are around 30 or 40.

The main drawback of IMS technology is that it provides moderate chemical selectivity. This can be a problem when multiple substances are present in a sample or in the presence of interferences as pollutants or moisture in environmental analysis (Márquez-Sillero et al. 2011). In these cases multiple peaks appear in the spectra and typically overlapped. Therefore, the identification and quantitation of substances of interest becomes problematic.
One way to improve selectivity in IMS is the use of doping agents which may be added to the drift gas in order to control ionization. For instance, in positive ion mode and no dopant, water, which has a low proton affinity, dominates the ionization process. This guarantees a response to any analyte with a proton affinity higher than water, but this can be problematic if interfering chemicals or humidity are present. When doping the instrument with an agent with a proton affinity higher than water, only those analytes with an even higher affinity will react, thus improving the selectivity of the instrument (Hill & Simpson 1997). A similar reasoning can be done for the negative ion mode. With this regard, explosives and some CWA are characterized by their high electronegativity, thus being IMS a suitable technique for their detection.
When using IMS instruments based on radioactive ionization sources (the most common configuration), the ionization of the sample occurs by ion/molecule reactions

rather than by direct ionization of the analyte; thereby, this kind of sources typically have a nonlinear and narrow dynamic response range (Hill & Simpson 1997), thus hindering qualitative and quantitative analysis. In IMS atmospheric chemistry (with no dopant), the formation of protonated molecules (MH^+) is due to an effect of proton transfer from the reactant ions, predominantly hydronium ions: $(H_2O)_nH_3O^+$, to the analyte molecules; this chemical process could occur either in the reactant region or in the drift region. Additionally, when the concentration of analyte molecules is adequately high, a proton-bound dimer is formed as a result of clustering of a protonated monomer with an additional analyte molecule (Jazan & Tabrizchi 2009; Eiceman & Karpas 2005).

(Eq. 1.5)   Protonated Monomer: $M + (H_2O)_n H_3O^+ \leftrightarrow MH^+ + (n+1)H_2O$

(Eq. 1.6)   Proton-Bound Dimer: $MH^+ + M \leftrightarrow MH^+M$
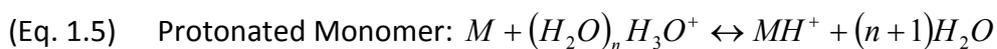
High concentrations also favour the formation of proton-bound trimers, proton-bound tetramers and so on (Young et al. 1999; Borsdorf, Stone & Eiceman 2005; Krylov et al. 2002; Ewing, Eiceman & Stone 1999), but normal IMS spectra do not show this ion species due to their extremely short lifetimes. In the presence of humidity, clusters with more water molecules can be formed, increasing the width of the peaks and thus reducing the resolving power (Eq. 1.3) and the peak-to-peak resolution (Eq. 1.4).

The effect of sample concentration on IMS response is shown in Figure 1.8; first the intensity of protonated monomer raises when concentration of the analyte is increased, and at the same time the intensity of reactant ion peak decreases. On the other hand, when a proton-bound dimer peak appears with a further concentration increase, the intensities of the protonated monomer and reactant ion peaks decrease. When the analyte is removed from the ion source, its concentration decreases and intensity declines for the proton-bound dimer and it increases for the protonated monomer peak. Similar patterns are observed in negative polarity. In IMS, this is a typical behaviour of a system that has RIP-monomer-dimer in equilibrium (Eiceman & Karpas 2005).
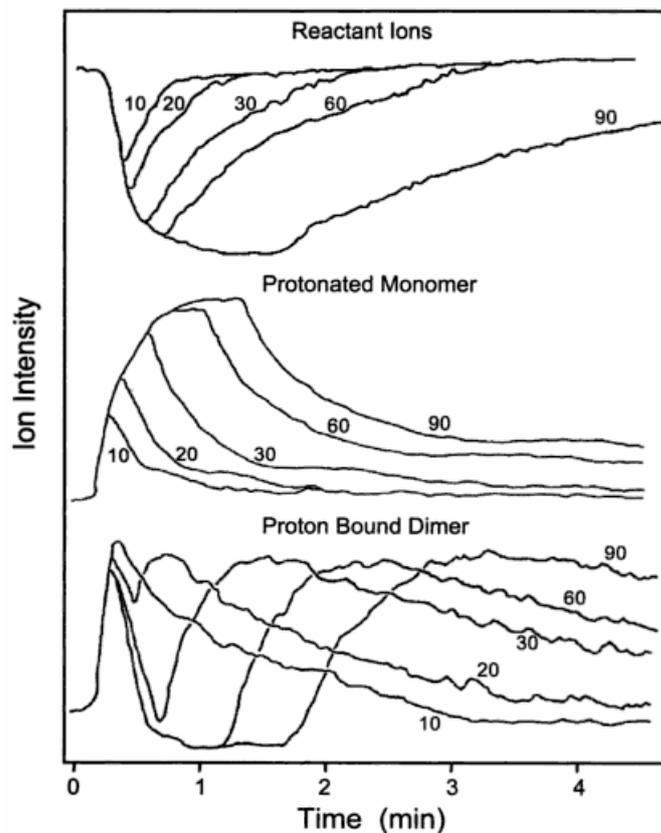
Figure 1.8. Plots of ion intensities from vapour sampling of a moving plume of a chemical vapour. The numbers are time in seconds used to generate the plume. These plots show the ion dynamics in an IMS analyzer. Extracted from (Eiceman & Karpas 2005).

Thereby, increasing the analyte concentration produces nonlinearities in the spectra. This effect worsens if more than one substance is present in the sample since nonlinear products among species can be formed, and it makes even worse in the presence of humidity (Puton et al. 2012).

Another way to improve selectivity and minimize nonlinear effects is coupling additional instrumentation to the IMS (Dworzanski et al. 1994; Kanu, Wu & Hill 2008). For instance, Gas chromotography (GC) (Simpson et al. 1996; Kanu & Hill Jr 2008), liquid chromatography (LC) (Garrido-Delgado et al. 2011) or multicapillary columns (MCC) (Ruzsanyi et al. 2005) can be used as a pre-separation step prior to IMS. Selectivity can also be enormously enhanced when using the IMS as a pre-filter before a mass spectrometer (MS) (Kanu et al. 2008; Sabo & Matejčík 2012).
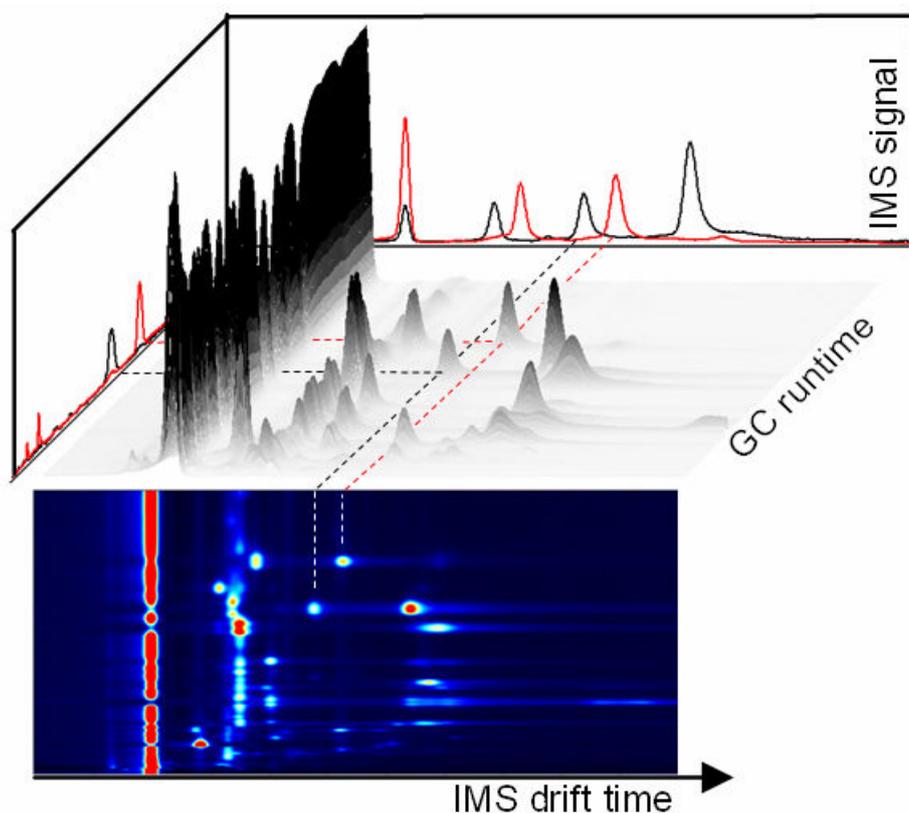
Figure 1.9. Example of signals obtained in (GC-IMS) (G.A.S, Dortmund, Germany).

In order to improve IMS selectivity, the issues of detection of substances in the presence of interfering chemicals and quantitation when dealing with non-linearities and multiple peaks in the spectra are addressed in the present thesis.

1.3.2 Differential Mobility Analyzer

Differential Mobility Analysis (DMA) is a particular configuration within the family of Ion Mobility Spectrometry where ions with different electrical mobilities are separated in space instead of drift time as in a classical drift-time IMS. This configuration should not be confused with DMS or FAIMS, where ions are also separated in space instead of in time, but based on the different mobility showed by the compounds when are under the alternative application of a low and a high electric field. DMS or FAIMS instruments are smaller than DMAs, the achievable resolving power (Eq. 1.3) is smaller and the results do not refer to real collision cross sections of the molecules and, therefore their size. High resolving powers imply better selectivities.

DMA has become the most common instrument widely used in the classification and generation of monodisperse aerosol nanoparticles; however, under certain conditions ionic detection can be achieved. The more used DMAs belong to the category of cylindrical and their evolution has been described in (Intra & Tippayawong 2008).

Cylindrical DMAs are based on Knutson and Whitby's design (Knutson & Whitby 1975) consisting in two cylindrical and concentric metal electrodes.

The DMA used in the present thesis has been developed by RAMEM S.A. (RAMEM) and corresponds to a parallel plate configuration (Santos et al. 2009; Alonso et al. 2009). The separation is done through the superposition of two perpendicular forces on the formed ion, one drag force given by a high sheath flow rate; and the second involved force comes from a perpendicular electric field. The planar ion-DMA scheme can be found in Figure 1.10, where the core of the instrument, the classification region, is shown.

A carrier gas containing ionized species enters the DMA through a slit and joins a particle-free sheath air, which flows between two parallel plate electrodes. Ions are dragged downstream the inlet slit by the sheath gas. Additionally, ions also migrate from one electrode to the other under the action of a uniform electric field, which is established by applying a voltage between the plates, that is, perpendicular to the direction of the sheath air. Only the ions of a given electrical mobility leave the DMA through a slit which is made in the outer electrode and collected by an ion plate connected to an electrometer. Changing the electric field between plates, ions with different mobilities are selected to exit through the slit where the detector is placed. Scanning the electric field, a spectrum in mobilities is obtained.

A similar configuration has been commercialised by Environics (Mikkeli, Finland) under the name of "aspiration type IMS" (ChemPro100). However, there are crucial differences between DMA and aspiration type IMS. The first difference is that the Environics' device uses a much smaller sheath flow, leading to much lower resolving powers (Eq. 1.3). The second difference deals with the detection, which is done in a multichannel configuration, with additional sensors (temperature, humidity and two semiconductors).

The classification in space using high sheath flow rates, instead of drift time have the advantage of achieving higher sensitivities since lower times are required to collect the charge from ions. Moreover, better resolving power might be achieved since shorter tubes can be used (Santos et al. 2009).

On the other hand, the resulting DMA instrument is much bigger and expensive than traditional IMSs. Measurement time is also longer although the assembly of a multi-collector can dismiss the needed time for the spectra recording.
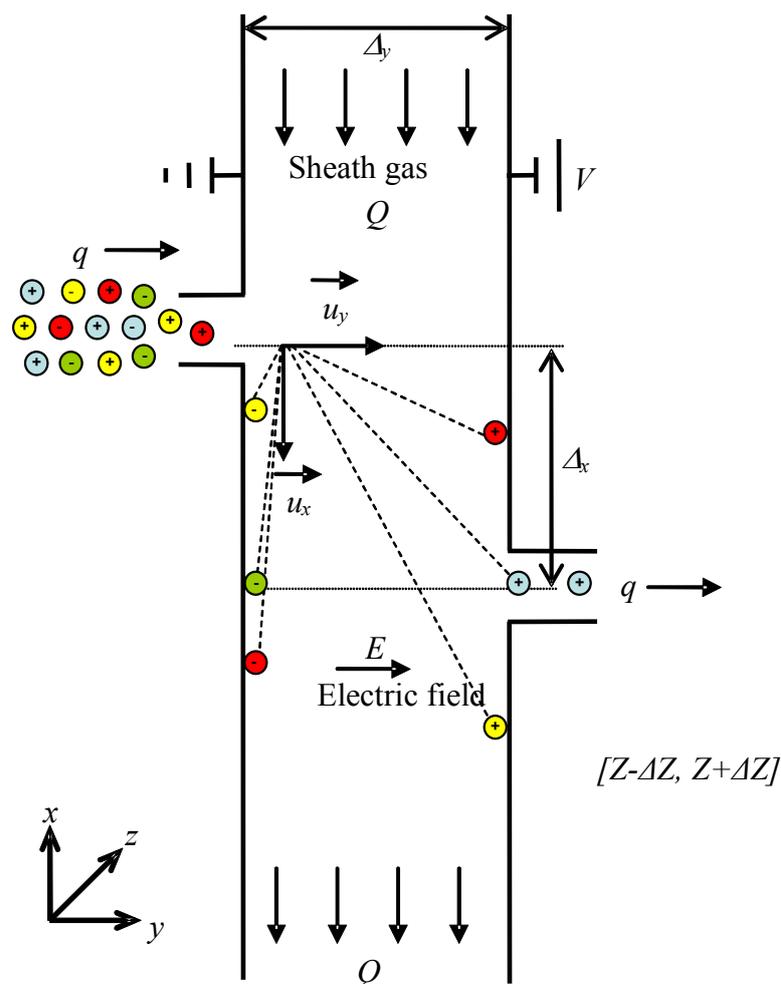
Figure 1.10. Sketch of a parallel–plate ion–DMA (RAMEM).

1.3.3 Other IMS configurations

Recently, new IMS configurations, apart from DMA, have shown up. For instance, Travelling Wave IMS (TWIMS) (Shvartsburg & Smith 2008; Giles, Williams & Campuzano 2011) which has parallels to both FAIMS (use of a periodic waveform) and drift time IMS (separation by absolute mobility). Transversal Modulation Ion Mobility Spectrometry (TM-IMS) has been presented in (Vidal-de-Miguel, Macía & Cuevas 2012) and claims to reach a RP of 50, although only one publication can be found so far. Trapped IMS (T-IMS) has been presented by (Fernandez-Lima et al. 2011; Fernandez-Lima, Kaplan & Park 2011) and also claim to reach RP of 80 to 120.

However, these IMS configurations are relatively new and how to process their signals still needs further work.

1.3.4 Photo Ionization Detector

A photo Ionization Detector (PID) is a field instrument that is relatively simple and easy to use. It provides rapid information (few seconds) about volatile organic compounds (VOCs) in air samples with high sensitivity (ppb levels). The critical component of a PID is a lamp, which produces photons in the ultraviolet (UV) energy range.

The sample is collected by a small air pump and introduced into the PID where it passes in front of the lamp and is exposed to the UV radiation. Atoms and molecules in the sample that have an ionization potential (IP) lower than the energy of the UV lamp are ionized with some efficiency. UV lamps are typically in the energy range from 8.4 to 11.7eV. An electric field then pulls ions to the appropriate electrode where a current can be measured. A diagram of a PID instrument is shown in Figure 1.11 (Daum et al. 2006).
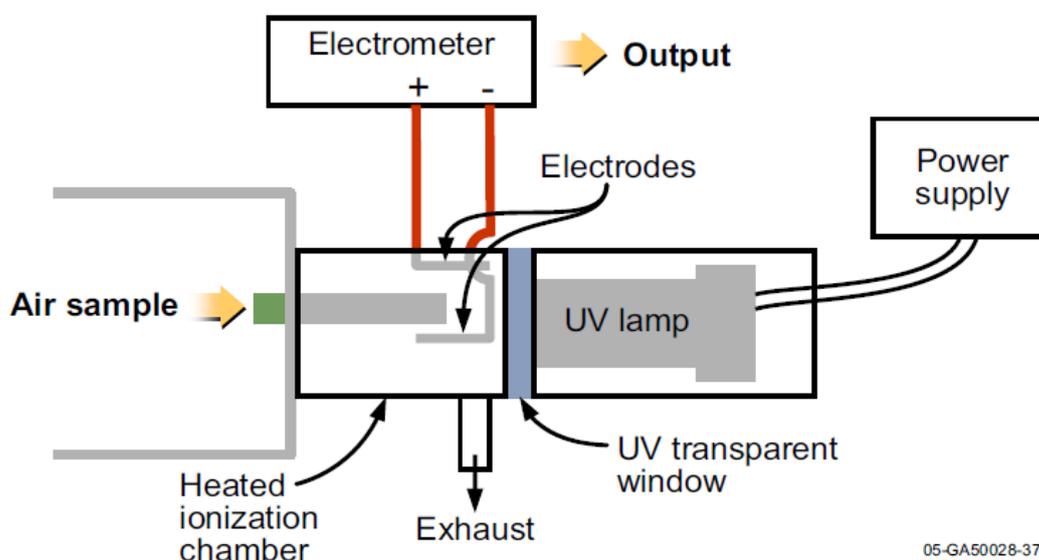
Figure 1.11. A PID instrument diagram (Daum et al. 2006).

A PID is non-selective in the sense that all molecules below the IP are ionized and producing a response, therefore it provides information about the relative magnitude of contamination but it is unable to distinguish specific compounds. Isobutylene is typically used to calibrate PIDs as it is stable, relatively easy to handle, readily available and can be stored at high pressure. Instrument responses for other gases are then obtained by multiplying the reading by a correction factor which takes into account the response relative to isobutylene.

Since the sensor is not selective to ionized compounds, PIDs can be used as a first screening tool to provide an initial warning; first responders can act and use a more specific sensor afterwards. The technology has been successfully applied to the

detection of CWAs, TICs and for environmental monitoring since the detector can respond to any chemical present in the air (Sferopoulos 2009).

Figure 1.12. Example of two portable commercially available PIDs. They incorporate an 10.6eV UV lamp allowing the detection of VOCs with a resolution of 0.1ppm and a time response below 3s. Top: (MiniRAE3000) with concentration range from 0.1ppm to 15000ppm. Bottom: (MultiRAEplus) which, apart from the PID, incorporates additional sensors (O_2 and selective sensors to typical toxic gases as CO, H_2S, SO_2, NO, NO_2, Cl_2, HCN, NH_3 or PH_3). (RAE systems, San Jose, USA).

1.3.5 Metal oxide sensors (MOX)

Metal oxide sensors (Meixner & Lampe 1996) are transducers that incorporate a chemical detection layer and transform a chemical interaction into an electrical signal. They are also usually capable of continuous measurements and are generally inexpensive.

The principle of operation is based on the change in the conductance (resistance) of the oxide in interaction with a gas (Zohora, Khan & Hundewale 2013). The gases, which act as reducing or oxidizing agents at MOX operating temperatures, take part in redox reactions on the surface of the MOX. The change in resistance is dependent upon the VOC that reacts with the adsorbed oxygen on the sensing surface, as well as the metal oxide grain size.

Selectivity can be gained by changing either the catalyst or the operating conditions for the sensor. MOXs generally function at 80-500ºC, allowing rapid and reversible reactions at the sensor surface and avoiding formation of a layer of chemisorbed water that would inhibit the reaction with VOCs.

Sensitivity is dependent upon the operating temperature, and careful temperature modulation has been shown to improve discrimination.

Commercial sensors have typical detection limits of 5-500ppm, though this depends on the analyte (James et al. 2005). The time response is typically less than five seconds, however the recovery time depends on the operating temperature. At low temperatures the recovery time is around one minute. Figure 1.13 shows a simple diagram of a MOX.
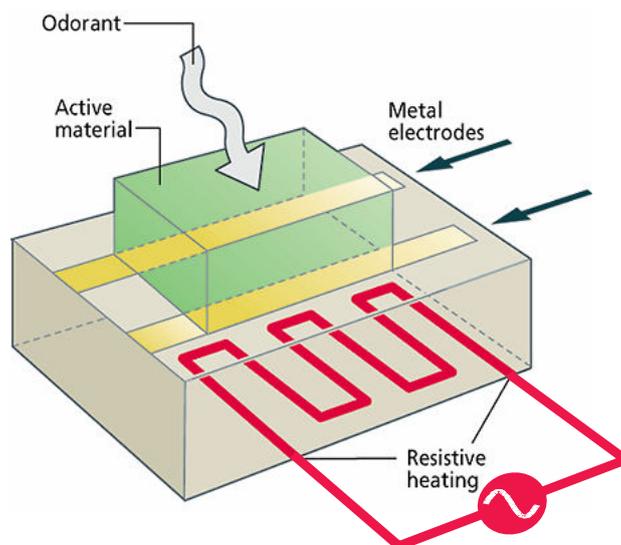
Figure 1.13. Basic MOX sensor diagram. (Nagle, Gutierrez-Osuna & Schiffman 1998).

Metal Oxide gas sensors are usually used in combination with other gas sensors forming arrays of partially selective sensors. Persaud and Dodd first reported the design of an array of reversible but only partially selective sensors (Persaud & Dodd 1982). This allowed increasing the selectivity of the system through the application of pattern recognition techniques to responses obtained from the sensor array (Scott, James & Ali 2006). They reported that the device, known later as an electronic nose (e-nose) (Gardner & Bartlett 1994), could reproducibly discriminate between a wide variety of odours.

Current sensor array technology devices are based upon the use of an array of several different chemical sensors such as conductive polymers, metal oxide, bulk acoustic wave and surface acoustic wave devices which can be used simultaneously for real time monitoring (Rock, Barsan & Weimar 2008; Vanneste & Geise 2004).

Multiple applications have been proposed for e-noses and gas sensor arrays. However, most of the research has focused on food industry (Schweizerberberich, Vaihinger & Gopel 1994), medical diagnosis (Gardner, Shin & Hines 2000), environmental monitoring (Bourgeois et al. 2003) and process control (Perera et al. 2006).

Continuous "in situ" monitoring of air, water and land quality is fundamental to most environmental applications. The ability and performance of gas sensor arrays under realistic conditions is discussed in (Bourgeois et al. 2003). One of the applications in the field of environmental monitoring is the detection of explosives (Yinon 2003) and toxic compounds (Mitzner, Sternhagen & Galipeau 2003). In (Hopkins & Lewis 2001) an array of composite polymer sensors is used for the detection of nerve agents in the presence of background VOCs such as benzene, methanol, toluene, lighter fluid, diesel fuel and tetrahydrafuran.

Some commercial electronic noses with multiple configurations can be found in the market. The border between classical analytical systems, electronic nose technology, and detectors for specific substances or even single compounds becomes more and more fuzzy. Some manufacturers call their devices "electronic noses", whereas others avoid mentioning this term even if their product operates in a similar way (Rock, Barsan & Weimar 2008).

Classical electronic noses which incorporate arrays of chemical sensors are manufactured by several companies, for instance Airsense Analytics (Schwerin, Germany), Alpha MOS (Toulouse, France), AppliedSensor (Reutlingen, Germany) or Smiths Detection (Watford, England), among others.
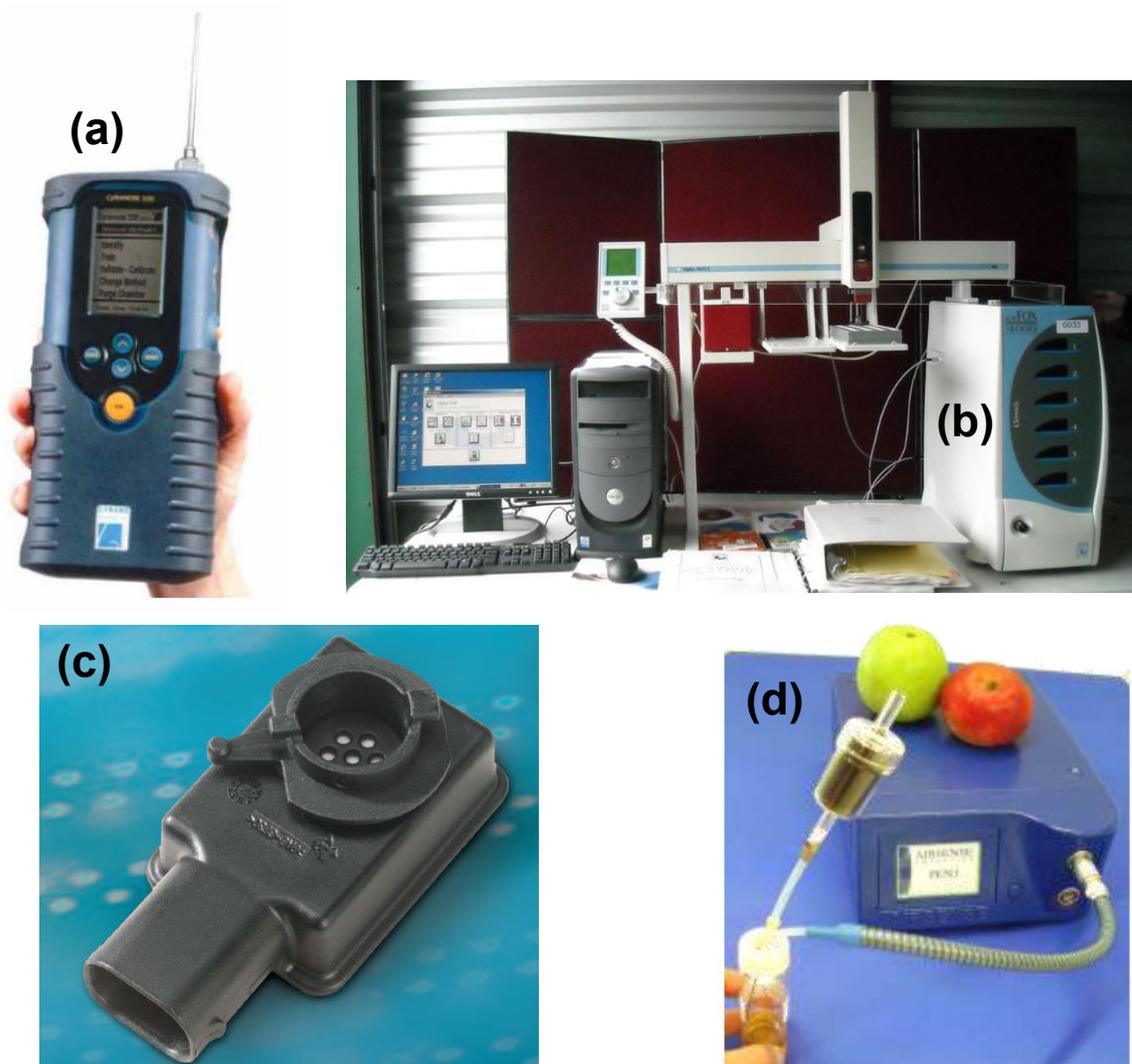
Figure 1.14. Commercial e-noses with classical configuration. (a) (Cyranose320) by Smiths Detection is a portable e-nose with 32 nanocomposite sensors. (b) (FOX4000) by Alpha MOS incorporates 18 MOXs mainly for VOCs detection. (c) (ACM) by AppliedSensor is an Air classification module for detecting traffic-related gases an preventing them for entering the automotive cabin. (d) (PEN3) by Airsense Analytics includes 10 different MOXs single thick film sensors.

The ACM Air Classification Module by AppliedSensor has been demonstrated to be highly successful (>100000 systems sold) to detect and quantify potentially harmful gases such as nitrogen oxides (NOx), carbon monoxide (CO) and volatile organic compounds (VOCs), providing fast responses (around one second), through the combination of MEMS (microelectromechanical systems) and signal processing, at low cost.

Additionally, configurations combining different sensing technologies in one instrument can also be found in the market (Figure 1.15).

Figure 1.15. Commercial instruments which include different sensing technologies. (a) (GDA2) by Airsense Analytics includes an ion mobility spectrometer (IMS), a PID, an electrochemical cell (EC) and 2 MOXs. (b) (RQbox) by Alpha MOS is a field based e-nose for odour and pollutant gases monitoring. It includes an EC, a PID and different MOXs.

1.4 Introduction to signal processing for IMS spectra (IMS & DMA)

Current instrumentation produces large amounts of data which need to be analyzed. Many times, depending on the application, it is not possible to analyze these data manually, for instance in continuous monitoring or real-time applications. Moreover, frequently, the relevant information provided by the instrumentation is hidden in complex signals or spectra. The detection and extraction of this relevant part in many cases is not trivial. With this regard, multivariate data analysis tools have been demonstrated useful for this purpose in numerous applications and diverse settings. While chemometrics and pattern recognition techniques have been largely applied in other instrumental techniques (Non-dispersive infrared sensors, e-noses, GC-MS, etc), the application to IMS techniques has been limited.

As pointed before (section 1.3.1), compared to gas chromatography, IMS[1] provides worse selectivity since wider peaks appear in the spectra, and this effect is aggravated by the presence of humidity, thus usually peaks appearing overlapped.

IMS can be considered as a technology related to electronic noses and gas sensor arrays, in the sense that the technique exhibits only a moderate selectivity. Instead of having an array of partially selective sensors with overlapped specifities, the response of the instrument corresponds to a spectrum of multiple points (drift times or mobilities) where adjacent points are correlated. Therefore, from the point of view of

---

[1] Henceforth, when using the term IMS, we refer to the general concept, including any particular configuration, as DMA for instance, belonging to the family of ion mobility based techniques.

data analysis, the response at each spectral point is interpreted as if it was the response of a particular sensor of a hypothetical array.

While limited selectivity is an issue, IMS spectra are plagued with other problems that can be corrected or at least partially compensated with signal and data processing. Spectra are often noisy (this is always the case when analytes are close to the detection limit), present a certain baseline which should be removed, and are sensitive to changes in humidity, pressure and temperature which makes the signal unstable, with peaks usually shifted leading to misalignments. Therefore, previously to applying multivariate techniques for analyzing IMS spectra qualitatively and quantitatively, a certain pre-processing of the signals is required.

1.4.1 Pre-processing IMS spectra

1.4.1.1 Filtering and baseline removal

Noise is a very common problem in chemical instrumentation which may lead to the deterioration of accuracy and precision, thus affecting negatively qualitative or quantitative results. Many methods, such as mean filtering, exponential smoothing, Savitzky–Golay (Savitzky & Golay 1964), and digital filtering in general, have been used to cope with noise problems. Recently, wavelet techniques have got more interest (Barclay, Bonner & Hamilton 1997; Shao, Leung & Chau 2003; Coombes et al. 2005). A formal interpretation of the term "de-noising" is given in (Donoho 1995), where it is shown how wavelets transforms may be used to optimally "de-noise" signals.

When removing noise from signals, baseline removal should be addressed specifically. While filtering of high frequency noise is a well established topic in the literature, the baseline problem (low frequency noise) is a relatively complicated one. Conventional frequency analysis cannot give a theoretic description of the baseline information, thus it is difficult to distinguish between baseline and other signals.

Approximate estimation of baseline has been the general method. When dealing with a single peak that requires the estimation of a local baseline, a straight line is used to connect the two ends of a signal peak. However, sometimes is difficult to set up clearly where the peak deviates from the baseline to select the first and last point of this local estimation of the baseline. The straight line is taken as the baseline and further calculation of peak area or peak height for substance quantitation is based on it. A more general method that can be used to correct the baseline of the full spectrum consists in fitting a polynomial of a certain order to the first and last interval of the

spectrum, where usually no relevant peaks appear. However, if the straight line or the polynomial does not fit the real baseline, the baseline estimation will lead to errors.

Recently, new approaches have been proposed to make a better estimate of the baseline. Although some parameters must be set initially, the main advantage of the new approaches is that signals are processed fully automatically through an iterative procedure (Komsta 2011).

For instance, in (Gan, Ruan & Mo 2006) the automatic thresholding method is used. Initially a polynomial of a certain order is used to fit the signal. All parts of the signal above the fitted line are cut to this line, and then iteratively new fittings are computed resulting in lines lying in lower and lower part of the signal. After several iterations, when a convergence criterion is reached, only the baseline remains, which then can be substracted from the signal.

The *airPLS* algorithm (available for R and MATLAB) is presented in (Zhang, Chen & Liang 2010). The method works by iteratively changing weights of sum squares errors (SSE) between the fitted baseline and original signals, and the weights of the SSE are obtained adaptively using the difference between the previously fitted baseline and the original signals.

An alternative approach based on asymmetric least squares is presented in (Peng et al. 2010). By means of the similarity among the multiple spectra, the algorithm estimates the baselines by penalizing the differences in the baseline corrected signals. In addition, a relaxation factor which measures the similarity of the baseline corrected spectra is incorporated into the optimization model and an alternate iteration strategy is used to solve the optimization problem.

Since these algorithms are quite general, they can be applied to different spectra and signals, including chromatograms, vibrational spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, Raman spectra and IMS spectra, for instance.

1.4.1.2 Peak alignment

Alignment of spectra is critical for analysis involving the comparison of multiple spectra. Analytical sensor data resulting from complex chemical mixtures can often be misaligned to time-varying biases resulting from instrument variability and shifts. In addition to the need for time alignment, data from IMS may be shifted when comparing various datasets due to the effect of temperature, pressure and humidity on the mobility of ions.

With this regard, the effect of variations in temperature and pressure can be compensated using (Eq. 1.2) when spectra are expressed as function of the ion mobility, which is inversely proportional to the drift time. However, this first order

correction may be not enough for some applications. Therefore additional techniques should be used before analyzing IMS spectra qualitatively and quantitatively.

Different algorithms have been proposed in the literature, for instance dynamic time warping (DTW) (Ramaker et al. 2003; Tomasi, van den Berg & Andersson 2004) or correlation optimized warping (COW) (Tomasi, van den Berg & Andersson 2004; Skov et al. 2006), which have been demonstrated to be effective on chromatographic data and also have been employed for solving simple NMR alignments with satisfactory results.

The Interval Correlation Optimised Shifting algorithm (*icoshift*) (Savorani, Tomasi & Engelsen 2010) has recently been introduced for the alignment of nuclear magnetic resonance spectra. The method is based on an insertion/deletion model to shift intervals of spectra/chromatograms and relies on an efficient Fast Fourier Transform based computation core that allows the alignment of large data sets in a few seconds on a standard personal computer. Computation time is significantly reduced compared to the COW algorithm. Moreover, *icoshift* proved to perform better than COW in terms of quality of the alignment (simplicity and peak factor), but without the need for computationally expensive optimisations of the warping meta-parameters required by COW. The method has been proved to be suitable for chromatographic data (Tomasi, Savorani & Engelsen 2011).

The alignment of IMS data usually requires a rigid shift in the drift time (or mobility) dimension due to linear changes in instrument detection (Krebs et al. 2006). This rigid shift ensures that spectra are not distorted, thus not losing their physical meaning. Since *icoshift* is based on intervals, the distortion in the signals is local and significantly reduced. Thereby, this method could be useful to align IMS spectra as it is shown in Figure 1.16.
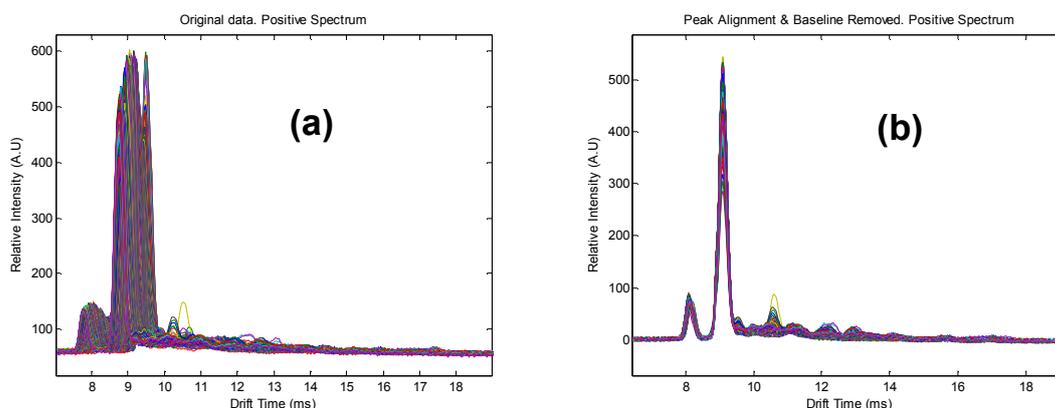
Figure 1.16. Example of pre-processing applied to IMS spectra. (a) Collection of IMS spectra with baseline and peaks shifted along measurement time due to changes in humidity and temperature. (b) IMS spectra with baseline removed through polynomial fitting and spectra aligned using the *icoshift* algorithm.

1.4.2 Multivariate data analysis

Once spectra have been pre-processed properly (filtering, baseline removal and alignment), multivariate signal processing methods can be applied in order to extract the relevant information needed for the specific application.

1.4.2.1 Dimensionality reduction and feature extraction

When working with large sensor arrays or spectrometers, since it is not uncommon to have hundreds or thousands of sensors or spectral points, the dimensionality of the data is large. However, the number of analysis that can be performed usually is scarce, thus resulting in a low number of samples. It is well-known that the performance of predictive (either qualitative or quantitative) models worsens when the ratio dimensionality/samples increases beyond a certain value. This is the well known curse of dimensionality (Bellman, 1961). Moreover, when using this kind of instrumentation, sensor measurements and spectra provide highly correlated and redundant information.

A dimensionality reduction step is useful for compressing and retaining the essential information from the data. After dimensionality reduction, data is more easily displayed and analyzed than each of the variables individually. Furthermore, essential information often lies not in any individual variable (sensor or spectral point) but in how the variables change with respect to one another; that is, how they co-vary. In this case, the information must be extracted capturing trends in the data.

Different algorithms have been proposed in the literature for dimensionality reduction and extraction of relevant information. Principal Component Analysis (PCA) (Jolliffe 2002; Wold, Esbensen & Geladi 1987) is probably the most used tool for dimensionality reduction and information extraction. PCA is an unsupervised exploratory analysis method which finds combinations of variables, or factors, which describe major trends where variance in the data is maximized. These factors are mutually orthogonal and form a new basis where data generally can be visualized more easily. It is found that data can be adequately described using far fewer factors than original variables with no significant loss of information. Additionally, PCA can find combinations of variables which are useful descriptions of particular events. Moreover, these combinations of variables are often more robust indicators than individual variables.

Multivariate curve resolution alternating least squares (MCR-ALS) (Tauler, Kowalski & Fleming 1993) is an iterative algorithm that provides a bilinear decomposition of the original data matrix and which converges to a local solution based on the minimization of the squared sum of residuals. While in PCA the loadings are orthogonal, in MCR

other constraints are introduced: typically non-negativity of the loadings and/or scores. This algorithm is useful to refine initial estimations (e.g. found using PCA) since knowledge of the system is incorporated within the iterative process, thus meaningful solutions are obtained. The reader is referred to section 1.4.2.3 for more details.

Unlike PCA and MCR-ALS which only use the responses from the instrument, other methods also incorporate information about the category or group each sample belongs to, also known as *class*.

Representing class information using a 1-of-K coding scheme is usually convenient. Given a class C_j, t is a vector of length K where all elements t_k are zero except element j, which takes value 1. Each sample has associated its vector t.

For instance, Linear Discriminant Analysis (LDA) (Fisher 1936) finds linear discriminant functions (LDF), which are linear combinations of the original variables, with the objective to maximize the ratio of the between-class scatter and the within-class scatter, thus maximizing class separability at the same time that attempts to minimize dispersion within each class.

Class information is also incorporated in Partial Least Squares Discriminant Analysis (PLS-DA). In (Barker & Rayens 2003) it is shown that PLS-DA is essentially the inverse-least-squares approach to LDA producing similar results but incorporating the PLS (see section 1.4.2.4) advantages of noise reduction and variable selection. This method tries to maximize the correlation between the instrumental responses and the classes.

Previous methods are based on linear functions of the input variables (instrumental responses). Although these methods have significant limitations particularly for problems involving input spaces of high dimensionality, they have nice analytical properties and form the basis for more sophisticated methods (Bishop 2006), such as nonlinear methods; for instance, artificial neural networks (ANN) and support vector machines (SVM).

1.4.2.2 Substance classification

The goal in classification is to take an input vector (instrumental response) and to assign it to one of K discrete classes C_k, where k is an index from 1 to K. In the most common scenario, the samples are assigned to only one class. The input space is thereby divided into decision regions whose boundaries are called *decision boundaries* or *decision surfaces* (Bishop 2006). These boundaries can be obtained from linear or nonlinear methods, including those which perform dimensionality reduction and feature extraction, such as LDA, PLS-DA, ANN and SVM.

Classification is a supervised method where samples obtained under known conditions (training dataset) are used to build a model. Using this model, unknown samples can be classified automatically.

The simplest measure of classifier performance is the classification rate, the percentage of samples assigned to the correct class.

The *k*-nearest neighbours (KNN) algorithm (Hastie, Tibshirani & Friedman 2003) is a simple and supervised technique to generate non-linear boundaries between classes in the training dataset. KNN finds the closest *k* samples to the new sample and assigns the predominant class to it. Typically the distance metric is either angular or Euclidean. The Euclidean distance (d_{ij}) is defined as:

(Eq. 1.7) $$ d_{ij} = \sqrt{\left(x_i - x_j\right)I_n \cdot \left(x_i - x_j\right)^T} $$

where *x* is a sample in a *n*-dimensional space, I_n is a *n*-dimensional unity matrix. The angular distance ($d_{\vartheta ij}$) is defined as:

(Eq. 1.8) $$ d_{\theta_{ij}} = \arccos\left(\frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|}\right) $$

The KNN algorithm is normally used after dimensionality reduction (PCA, LDA, PLS-DA), thus applying the classifier over the reduced space.

1.4.2.3 Blind source separation

Multivariate Curve Resolution (MCR) techniques, belonging more generally to the class of algorithms known as Blind Source Separation (BSS) techniques in Signal Processing (Cichoki et al. 2009), aim to recover the evolution of the source signals (in our case, concentration profiles) and the mixing matrix (spectral features) without any prior supervised calibration step.

While BSS techniques (Amari, Cichocki & Yang 1996; Bell & Sejnowski 1995; Hyvärinen & Oja 2000) are popular in other domains (Young 2005; Vigário et al. 1998; Vigário 1997; Hyvärinen 1999; Back & Weigend 1997), their application for embedded intelligence in chemical instrumentation is still limited. However, in some conditions, basically linearity, BSS may fully recover the concentration time evolution and the pure spectra with few underlying hypothesis. This is extremely helpful in conditions where non-expected chemical interferences may appear, or unwanted perturbations may pollute the spectra. Moreover, in real world scenarios, spectra may become messy

featuring overlapping peaks and bad signal-to-noise ratio (Hill & Simpson 1997). In these conditions, BSS techniques are an option for spectra deconvolution and exploratory data analysis.

BSS methods are typically based on the assumption that the observed signals are linear superpositions of underlying hidden source signals, thus spectra and concentration profiles are obtained under the assumption of linearity:

(Eq. 1.9) $D = C \cdot S^T + E$

D is the data matrix (dimensions M x N) containing the observed signals from the instrument. Each row is a different spectrum/observation and each column is a different variable/sensor/component. C has dimensions M x K and contains the concentration profiles (concentration vs. time) related to the K pure components. S has dimensions N x K and contains the spectra associated with each pure component. E is a matrix of residuals (dimensions M x N).

We observe the data matrix D and we are interested in obtaining C, that is, the evolution in time of the underlying components present in the sample. BSS techniques provide C without any prior knowledge about S. That means C is obtained without any prior knowledge about the composition of the mixture. BSS are unsupervised techniques that give us C and S, apart from a scale factor, and no calibration is needed.

While techniques such as PCA calculate factors based on mathematical properties such as capturing maximum variance and orthogonality, the factors are often difficult to interpret because they are generally not directly related to the chemical process. For example, PCA loadings of a data set of measured spectra generally are not pure component spectra. Instead, the loadings are typically linear combinations of pure analyte spectra that have positive and negative intensities.

In other BSS methods in the signal processing domain, the source signals (pure chemicals in the present work) are typically assumed to be statistically independent. This is the case for Independent Component Analysis (ICA) (Hyvärinen & Oja 2000). ICA is a popular technique that has been successfully applied in a large number of applications (Young 2005; Vigário 1997; Vigário et al. 1998; Hyvärinen 1999; Back & Weigend 1997): separation of artifacts in Magnetoencephalography (MEG) data, finding hidden factors in financial data, noise reduction in images or telecommunications. However, it can not be applied to IMS spectra since the assumption of statistical independence among the source signals is not fulfilled in general. In many occasions, reactions taking place among the analytes (either inside or

outside the instrument), produce correlated time evolutions that violate the assumption of statistical independence. Therefore, alternative BSS techniques should be applied to IMS spectra.

A number of proposals have been originated in the chemometrics field. For instance, SIMPLISMA (SIMPLe-to-use-Interactive Self-modeling Mixture Analysis), proposed by Windig (Windig & Guilment 1991), has been applied to IMS data previously (Buxton & Harrington 2001; Cao, Harrington & Liu 2005). The algorithm looks for vectors maximally decorrelated and maximally pure. A pure variable is defined as that which has only contributions from one of the components in the sample. A geometrical interpretation of the purity concept is given in Figure 1.17 and in the original paper.
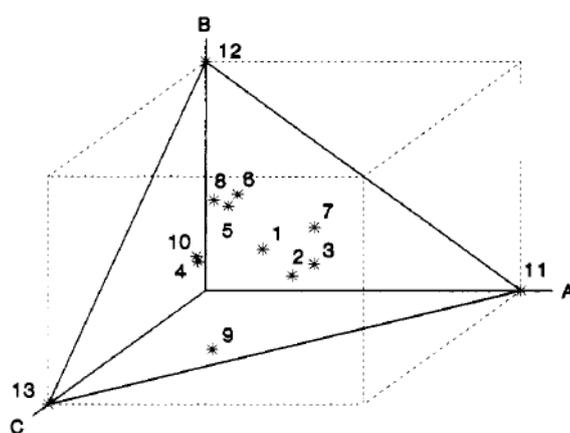
Figure 1.17. Three-dimensional plot containing thirteen mixtures of up to three components. Samples 11, 12 and 13 contain pure substances. Due to the fact that the concentrations of the three components add up to one, all the points representing the mixtures lie in a plane. Extracted from (Windig & Guilment 1991).

Another widely used algorithm is MCR-ALS proposed by Tauler (Tauler, Kowalski & Fleming 1993). An abundant literature exists on the use of MCR-ALS for resolving linear second order data and the algorithm has been applied to spectrophotometric (Azzouz & Tauler 2008; Jaumot et al. 2005), voltammetric (Antunes et al. 2002; Esteban et al. 2000) or chromatographic datasets (Tauler 1995). However, as far as we know, the use of MCR-ALS to resolve IMS datasets has been scarcely reported (Chen 2003; Cao & Harrington 2004; Buxton & Harrington 2001; Cao, Harrington & Liu 2005; Lu, O'Donnell & Harrington 2009; Chen 2008). Only a few works combining IMS and MCR have been published: regarding explosives detection (Buxton & Harrington 2001; Cao, Harrington & Liu 2005), drug detection (Lu, O'Donnell & Harrington 2009), determination of active principal ingredients (APIs) at low concentrations in pharmaceuticals (Zamora & Blanco 2012; Armenta & Blanco 2012; Armenta & Blanco 2011) and improved classification (Chen & Harrington 2008).

For instance, in (Armenta & Blanco 2011) it is shown the combination of IMS with MCR-ALS as an effective methodology for the detection of benzodiazepines in binary, ternary and higher order mixtures when peaks appear overlapped.

In contrast to SIMPLISMA, a single-step algorithm, MCR-ALS is an iterative algorithm which requires initial estimations. Many different strategies could be used to obtain first estimations such as: singular value descomposition (SVD) (Golub & Reinsch 1970), PCA, evolved factor analysis (EFA) (Keller & Massart 1992) or SIMPLISMA (Windig & Guilment 1991; Windig et al. 2005). The number of components K can also be selected using SVD and visual inspection of C and S obtained from SIMPLISMA.

For a given dataset, the solution of the bilinear decomposition (Eq. 1.9) is not unique and there is an infinite number of solutions for C and S due to the rotational (Jaumot & Tauler 2010) and scale (Tauler, Smilde & Kowalski 1995) ambiguities.

Scale ambiguity is not a serious problem in qualitative identification and an arbitrary normalization is usually applied either to C or S. In order to minimize rotational ambiguity and reduce the space of possible solutions, some constraints must be introduced (Gemperline & Cash 2003; Juan et al. 2000; Juan et al. 1997). This also results in an improvement in selectivity and thus in the interpretation of the results.

Basic constraints are: non-negativity, unimodality, closure, selectivity or local rank. These kinds of constraints are also called natural constraints since they frequently have a physical and chemical meaning and are frequently fulfilled by natural systems. Hard constraints modeling underlying physico-chemical processes (Juan et al. 2000) or the spectra shape can also be imposed in the solution.

The algorithm assumes that experimental data follow a bilinear model consisting of the decomposition of a dataset into the product of two sub-matrices of reduced sizes. In matrix form, this model can be written as in (Eq. 1.9). Thus, from initial estimations, MCR-ALS optimizes and solves the concentration profiles and spectra matrices iteratively through two least-squares steps:

(Eq. 1.10)
$$S^T = C^+ D$$

(Eq. 1.11)
$$C = D\left(S^T\right)^+$$

Where C^+ and $\left(S^T\right)^+$ are the pseudoinverse of the concentration and spectra matrix respectively. Figure 1.18 shows the MCR-ALS scheme used in this thesis. Initially, the original data matrix is filtered using principal component analysis (PCA) with K

components. Then, the ALS loop starts with initial estimations for spectra or concentrations profiles.

In each iterative step, different constraints can be used within the main ALS loop after computing (Eq. 1.10) or (Eq. 1.11). Convergence of the algorithm is assessed calculating the root mean squared error (RMSE):

(Eq. 1.12)
$$RMSE = \sqrt{\frac{\sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}\left(D_{m,n} - \hat{D}_{m,n}\right)^2}{M \cdot N}}$$

where $D_{m,n}$ represents the elements in the experimental data matrix D, and $\hat{D}_{m,n} = \left(C \cdot S^T\right)_{m,n}$ represents values computed by using concentration profiles and spectra obtained from the MCR-ALS method. The algorithm stops when the relative differences between RMSE for successive iterations is small enough (0.1%) or the maximum number of iterations is achieved. Typically the algorithm converges to local minima depending on the initial conditions.
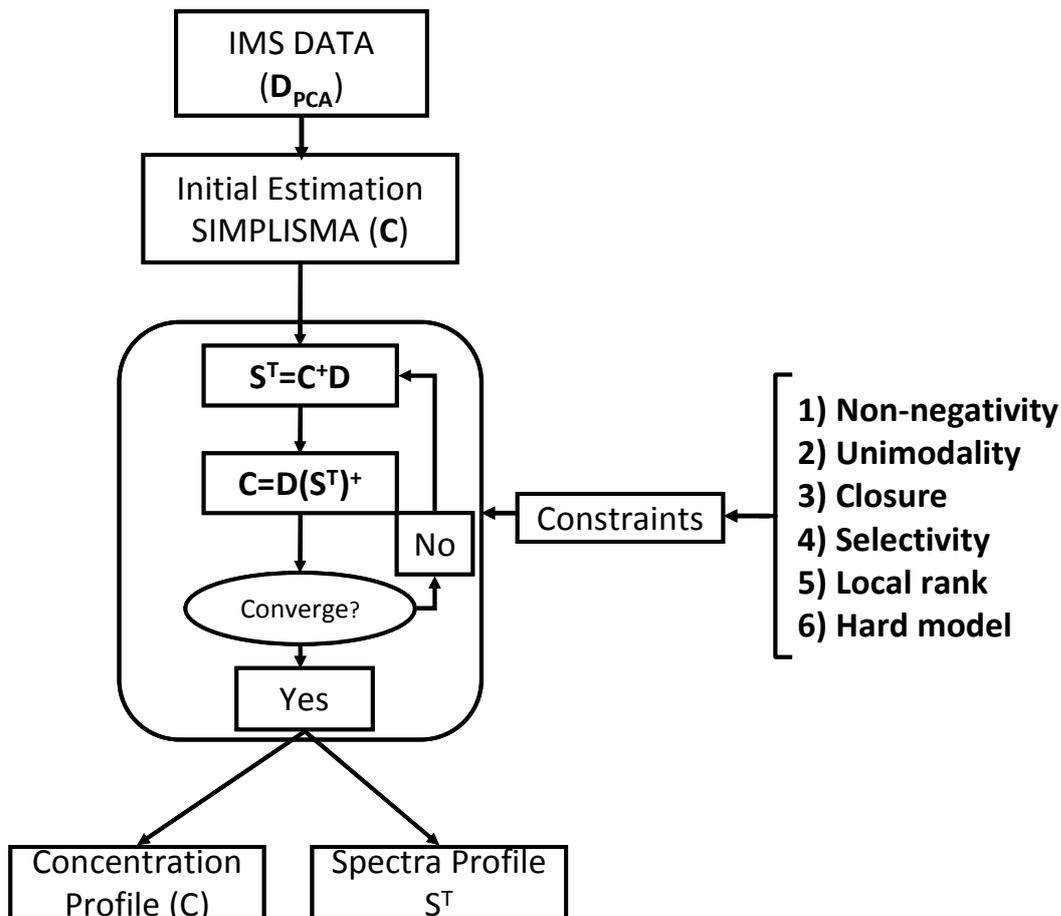
Figure 1.18. Block diagram of the MCR-ALS procedure applied to IMS spectra.

When having little knowledge about a mixture, it is seen that different multivariate techniques can be applied in order to obtain a qualitative estimation of the instrumental response (in particular IMS).

1.4.2.4 Quantitative prediction

Similar to the classification problem, quantitative prediction of samples is a supervised problem. Samples obtained under known conditions are required in order to build a calibration model, but instead of assigning a class to new samples, a quantitative value is assigned, which usually is the concentration of the substance.

In most of the applications, IMS or DMA instruments are used only as qualitative detectors and only the absence or presence of the target substance is of interest. However, in some other applications in certain scenarios, substance quantitation can become very important.

Quantitative determination of analytes with IMS is typically univariate. A calibration could be performed using the information of peak area or height and then applying an appropriate fitting function, which typically is polynomial.

However, such techniques are not useful if peaks are overlapped (unless a previous deconvolution is done). While it is possible to keep the instrument in a linear regime for low input concentrations, this seriously damages the dynamic range (ratio of maximum concentration to the limit of detection) of the instrument for most analytes. Moreover, if peaks behaviour is non-linear as concentration is increased, and more than one peak is related to the same substance, quantitation is degraded if only one of the peaks is used. As it can be seen in Figure 1.8, a calibration model based on the protonated monomer is very sensitive at low concentrations and provides no information at high concentrations; on the other hand, a calibration model based on the protonated-bound dimer is quite the opposite. Therefore, multivariate calibration techniques appear to be a good choice dealing with non-linearities and peak overlapping.

Different general purpose multivariate calibration techniques have been proposed in the literature: multiple linear regression (MLR) (Wise et al. 2006), principal component regression (PCR) (Vigneau et al. 1997) or partial least squares (PLS) (Wold, Sjöström & Eriksson 2001; Stone & Brooks 1990), among others.

These techniques are based on a linear relationship between the matrix *X* of spectral responses and the matrix *Y* of concentrations:

(Eq. 1.13) $Y = X \cdot B$

being *B* the matrix of regression coefficients which are obtained in the calibration step. Although the relationship between *X* and *Y* is linear, non-linearities can be modelled as well. The methodology to obtain *B* depends on the technique into consideration.

MLR, also known as ordinary least squares (OLS), is an easy way to perform a multivariate calibration; however, in spectroscopy, usually the number of samples is less than the number of variables and highly collinear variables are often found. In such conditions MLR is very sensitive to overfitting, which means that the calibration model may fit very well data but fails when used to predict the properties of new samples.

PCR is one way to deal with the problems of MLR. Instead of regressing the properties of interest (concentrations) onto a set of the original response variables (spectral response variables), the properties are regressed onto the principal component scores of the measured variables which, by definition, are orthogonal and therefore well-conditioned. Because it directly addresses the collinearity problem, PCR is less sensitive to overfitting than MLR, but one could overfit if the number of components to retain in the model is too high. In this sense, the PCR solution converges to the MLR solution as more components are added to the calibration model.

PLS regression is related to both MLR and PCR. PLS attempts to find factors which both capture variance (MLR case) of the spectral responses and achieve correlation (PCR case) between the spectral responses and the magnitude to be predicted (concentrations). Nevertheless, since loadings (in PCR) or latent variables (in PLS) are obtained from linear combinations of the original variables, the interpretation of the regression coefficients is usually difficult in such techniques. Moreover, the regression coefficients can have negative values, which have no sense from a physical or chemical point of view, and leading to compensation effects which increase even more the difficulty of physical or chemical interpretation. Furthermore, despite the fact that PLS algorithm is able to handle slightly non-linear data by increasing the number of latent variables in the calibration model, this approach is less successful for datasets containing moderate and severe non-linearities (Yang, Griffiths & Tate 2003).

Several variants of linear PLS have been developed so as to deal with non-linear datasets (Rosipal 2008), for instance using splines (Wold 1992) or using a polynomial of a certain order (Wold, Kettaneh-Wold & Skagerberg 1989; Baffi, Martin & Morris 1999).

However, latent variables remain difficult to interpret and for a physical and chemical interpretation of the process, MCR-ALS should be used instead as pointed in section 1.4.2.3.

Multivariate calibration has been applied to IMS only scarcely (Zheng, Harrington & Davis 1996; Fraga, Kerr & Atkinson 2009; Ochoa & Harrington 2005; Zamora & Blanco 2012; Lu, O'Donnell & Harrington 2009).

Recently, (Fraga, Kerr & Atkinson 2009) compared the performance of PLS and PCR with univariate regression based on peak area in the quantification of TNT and RDX in explosives. It is shown that multivariate calibration methods can be applied directly for quantitative prediction in IMS spectra and they provide better IMS quantitative precision and accuracy than univariate methods even when the peaks are resolved.

Also recently, Zamora et al. (Zamora & Blanco 2012) showed how two active principal ingredients (API) at low concentrations measured with IMS were also successfully quantified using PLS models, although peaks appeared overlapped.

1.5 Introduction to source localization algorithms

The environments where chemicals could propagate are very diverse and include underwater, above ground, below ground and enclosed spaces. Moreover, the dispersion of these chemicals in real scenarios is usually a very complex process where different factors participate, mainly: atmospheric stability, fluid flow direction and velocity, temperature, humidity and the presence of obstacles. Therefore, it is almost impossible to design a general source localization strategy capable to perform efficiently under all conditions.

Odour source localization is the act of finding the location of a volatile chemical source in the environment (Kowadlo & Russell 2008). Solving the source localization problem could be very useful for a large number of potential applications, often humanitarian: finding the source of dangerous substances such as airborne biological material and hazardous chemicals in industrial and other settings; detecting materials such as plant matter and drugs in a customs or quarantine application; searching for survivors in earthquake-damaged buildings, landslides or avalanches; detecting fire in its initial stages; locating unexploded mines and bombs; or detection and localization of clandestine drug/explosive laboratories, among many others.

A great number of approaches have been proposed in the literature trying to solve the source localization problem. Being these approaches implemented in both simulated

and practical experiments, with a wide variety of platforms, in several environments and for different applications.

Some reviews have also been published in order to group the different algorithms in categories, which is really useful so as to have an overview of the current state-of-the-art methods.

The basic issues regarding gas source localization includes the tasks of gas finding, gas tracking and source declaration. However, some strategies deal with gas distribution mapping rather than gas tracking. In the review published by Lilienthal et al. (Lilienthal, Loutfi & Duckett 2006) all these issues are covered, focusing mainly on experimental work and airbone chemical sensing. The paper presents a detailed summary of different methods and emphasizes the difficulty of the source localization problem in a real environment.

The different source localization strategies could be grouped based on different criteria, but probably the most comprehensive classification provided in the field is the one published by Kowadlo and Russell (Kowadlo & Russell 2008). In this paper, the different methods are grouped depending on the environmental conditions (fluid flow) and the localization method (taxonomic classifications). This provides a framework in which to evaluate the methods and compare them qualitatively. The article presents a survey of the existing approaches (until 2008) implemented in both simulation and practical experiments.

The more recent review of methods is the one published by Ishida et al. (Ishida, Wada & Matsukura 2012). It provides a brief history, the current trends of research and an overview of the existing methods for localizing chemical sources including useful references. Moreover, future directions and open issues are commented in the paper.

With this regard, little attention has been paid to the problem of localizing multiple time-varying sources, which can be of different kinds (physical, chemical, biological, electromagnetic), using multiple mobile robots (swarms of robots) and avoiding obstacles and collisions with other robots. In (Mcgill & Taylor 2011) general solutions are provided to address these issues; however, the results come mainly from simulations and, as it is stated in the paper, no common set of validation cases and reference algorithms are available for comparative analysis.

Figure 1.19. Venn diagram of the reported odour localization approaches until 2008. (Kowadlo & Russell 2008).

In the present section the difficulties associated to odour dispersion in a real environment are stated and the existing approaches are summarized taking as a basis the classifications provided by (Kowadlo & Russell 2008) but also considering the other reviews in the field (Lilienthal, Loutfi & Duckett 2006), (Mcgill & Taylor 2011) and

(Ishida, Wada & Matsukura 2012). At the end of the section, open issues and directions for future work are pointed.

1.5.1 Odour dispersal

The success in solving the source localization problem depends on how the given algorithm is suited to the environmental conditions, which determine how the odour is dispersed. The Reynolds number provides a measure of the level of turbulence in a fluid flow.

At low Reynolds numbers dispersion is dominated by viscosity and odours are dispersed by diffusion creating smooth concentration gradients. The maximum concentration is found at the source location and it decreases following a Gaussian distribution as the distance to the source increases (Hinze 1975). These are the conditions encountered by bacteria and where there is no fluid flow.

At medium to high Reynolds numbers, dispersion is dominated by turbulence. In these cases, odour dispersal is considerably faster than molecular diffusion, creating an odour plume downwind from the source (Figure 1.20). Under turbulent and time-varying flows, plume can meander and creates patches of odour gas, thus appearing intermittent regions with high concentration (much higher than the average) with steep gradients at their edges. Moreover, instantaneous concentration gradients fluctuate in intensity and direction (Figure 1.21 and Figure 1.22). These are the conditions which we should expect in a real environment where robots are deployed so as to localize gas chemical sources.
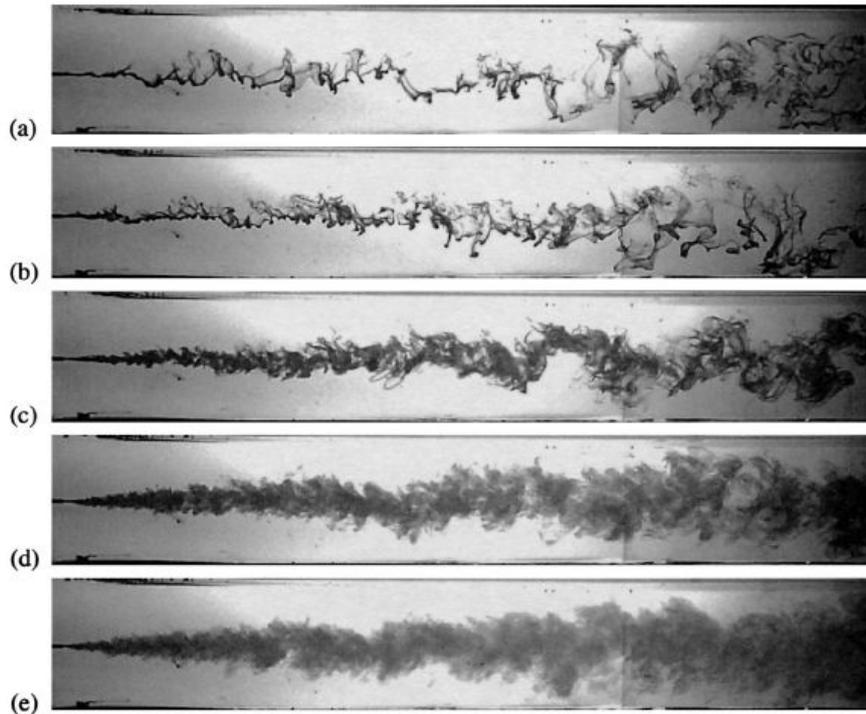
Figure 1.20. Plume released at different fluid flows with Reynolds numbers based on the relative velocity to the flow. (a) 0, (b) 227, (c) 454, (d) 908, (e) 1816. (Webster & Weissburg 2001).
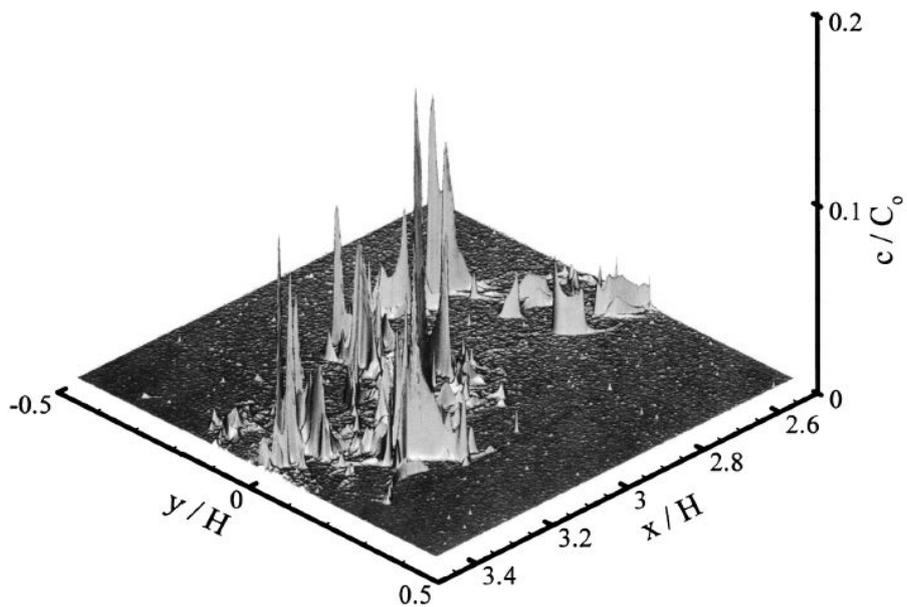
Figure 1.21. Example of instantaneous concentration fluctuations encountered in a turbulent fluid flow at different distances downwind (*x*) and crosswind (*y*) from the source. Concentrations are normalized to the source intensity (*C₀*) and distances are normalized to the flow depth (*H*). (Webster & Weissburg 2001).
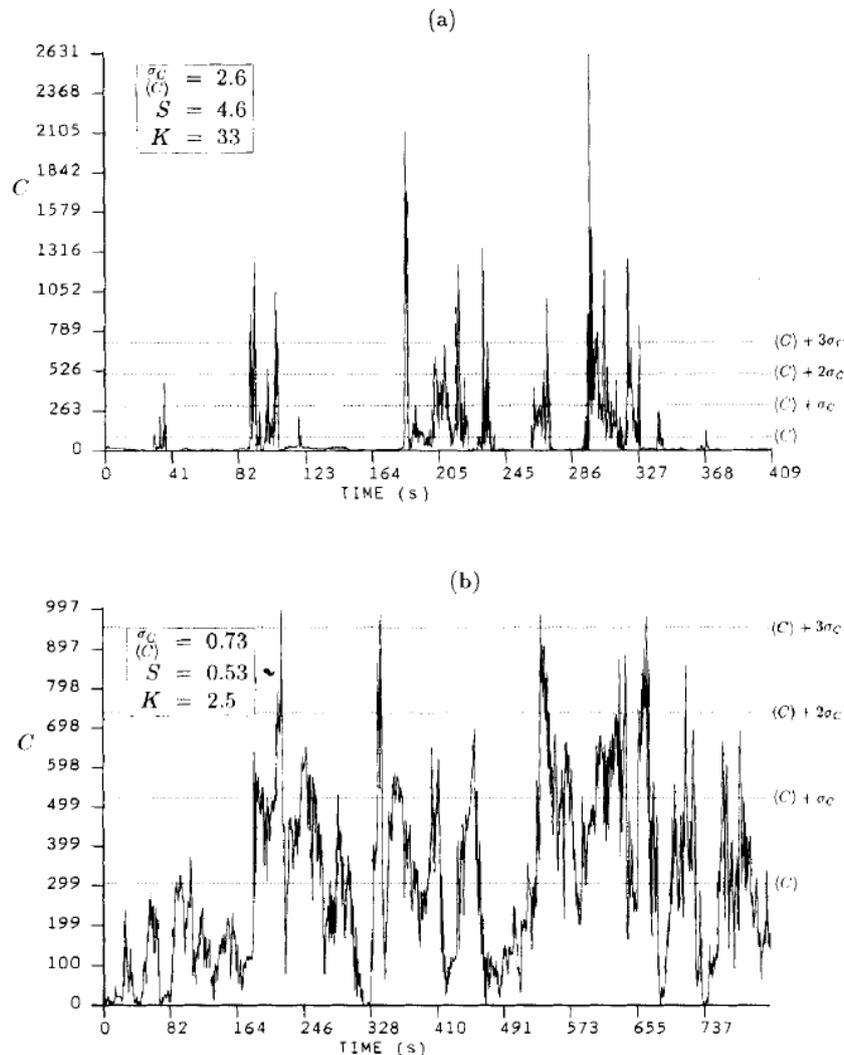
Figure 1.22. Example of sample concentration time series, comprising quantized values at 10Hz from a detector at a fixed point downwind of a continuously emitting source. (a) 80m from the source at a point off the plume centerline (y/σ_y=1). (b). 620m from the source and in the plume centerline (y/σ_y=0). In the graph it is shown, the mean concentration ($<C>$) and standard deviation σ_C of each time series, and the fluctuation intensity ($\sigma_C/<C>$), skewness (S) and kurtosis (K). Extracted from (Mylne & Mason 1991).

Although in a turbulence dominated flow the instantaneous characteristics of the plume are patchy, previous works have demonstrated that the time-averaged plume concentration follows a Gaussian distribution across the flow direction (Sutton 1947; Fackrell & Robins 1982; Crimaldi, Wiley & Koseff 2002; Webster & Weissburg 2001). This model has been widely used for its simplicity and is appropriate when dispersion is governed by atmospheric turbulence. Atmospheric turbulence is determined by the stability of the atmosphere and the height above the surface layer (Bakkum & Duijm 2005). The basic expression for the Gaussian Plume Model (GPM) for a continuous release is:

45

(Eq. 1.14)
$$C(x,y,z) = \frac{q}{2\pi U_a \sigma_y \sigma_z} \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \exp\left(-\frac{(h-z)^2}{2\sigma_z^2}\right)$$

where C is the mean concentration in g/m^3 in a location with coordinates x (downwind), y (crosswind) and z (vertical); q is the continuous source release rate in g/s; U_a is the mean wind speed in the downwind direction in m/s; h is the plume height in metres; and σ_y, σ_z are the diffusion coefficients (in metres) modelled as: $\sigma_y = a \cdot x^b$, $\sigma_z = c \cdot x^d$, where a, b, c and d are parameters obtained from a table (Bakkum & Duijm 2005) and their values depend on the atmospheric conditions which can be organized in six levels (from A-very unstable to F-very stable). In (Eq. 1.14), the resulting concentration distribution is due to the transport of chemicals by advection (due to the mean wind speed) and due to concentration gradients within the plume width (lateral dispersion due to diffusion, but also turbulent mixing). The decay of mean concentration is exponential, thus concentration levels below the sensor detection limit are achieved very fast.
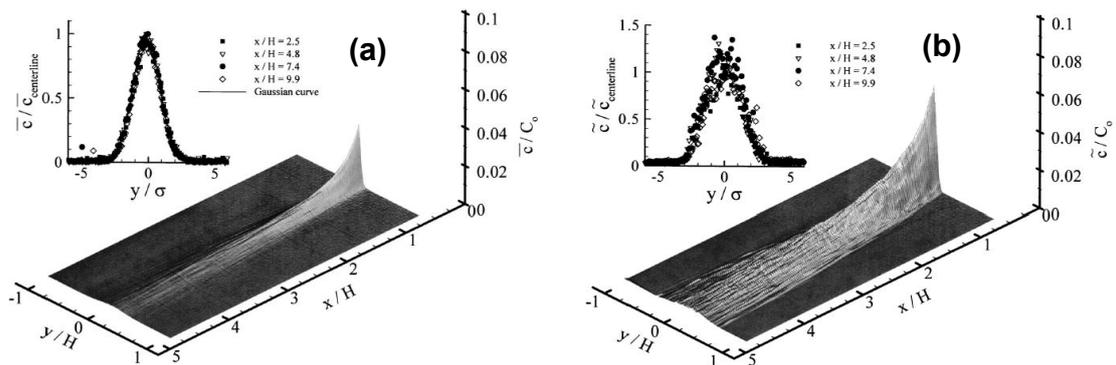
Figure 1.23. Mean plume characteristics in a turbulent flow at different distances downwind (x) and crosswind (y) from the source. Concentrations are normalized to the source intensity (C_0) and distances are normalized to the flow depth (H). (a) Time-averaged concentration field. (b) Standard deviation of the concentration field. Adapted from (Webster & Weissburg 2001).

The GPM only takes into account the time-averaged characteristics of a plume dispersed in a turbulent flow, but the sensors will be navigating and sensing the instantaneous plume characteristics. In this thesis, we assume that the sensor time response is much faster than the typical 10 min time-averages considered in the GPM. For short-time-scale studies, the chemical filament movement can be modelled as a random walk (due to velocity fluctuation) superimposed on the downflow advection (due to mean velocity); this is the idea published in (Farrell et al. 2002). An alternative idea to model the unpredictable and random fluctuations in concentration due to turbulent stirring and plume meandering is to consider the works by Eugene Yee et al.

(Yee, Wang & Lien 2009; Yee 2009; Yee 2008; Yee & Biltoft 2004; Yee & Chan 1997). They have carried out empirical studies on plume statistics in urban areas using a scale fluid model in a variety of plume conditions and urban geometries. Their results prove that instantaneous concentration fluctuations fit very well the clipped-gamma probability density function (PDF) over a very wide range of atmospheric conditions (from moderately convective to extremely stable stratification) and with receptor positions in both lateral and vertical cross-sections through the plume at downwind distances between 12.5m to 1km from the source (Yee & Chan 1997). The reader is referred to chapter 6 for the details.

1.5.2 Diffusion dominated fluid flow

In initial works in the field of odour localization in the 1990s, it was assumed that odour dispersion was dominated by diffusion creating a smooth chemical gradient which was used by a robot to move towards the source. This searching strategy is known as "chemotaxis" (Bell & Tobin 1982; Louis et al. 2008). The first implementation in a mobile robot was done by (Rozas, Morales & Vega 1991), and with swarm simulations by (Genovese et al. 1992), (Sandini, Lucarini & Varoli 1993) and (Buscemi & Prati 1994). Although these works were pioneering, opening a new field of study, the assumption of a chemical gradient is inaccurate in airborne chemical sensing, since at least a weak fluid flow is always expected to be found. However, in certain scenarios, for instance underground, the chemical gradient assumption is valid. Finding a chemical source underground has important applications, as identifying gas/fuel leaks from pipes and storages facilities, finding unexploded mines, victims of avalanches or earthquakes. With this regard, algorithms have been published to localize odours underground using a buried probe moved by a robotic manipulator in two (Russell 2004b) and three dimensions (Russell 2004a; Russell 2011).

Other works based on a concentration gradient, but using sensor networks have been published recently (Fiorelli, Bhatta & Leonard 2003; Zhang, Sobelman & He 2009). The ability to find multiple sources is also tested (Borah & Balagopal 2004).

1.5.3 Turbulence dominated fluid flow

The fluid flow could be affected by forced ventilation or convections due to temperature gradients; moreover, the presence of obstacles could complicate the plume dispersion.

Although chemotaxis was initially defined as "orientation or movement of an organism in relation to a chemical gradient in a diffusion dominated flow", more recent works

47

have extend the definition to "orientation or movement in relation to chemical agents" (Kowadlo & Russell 2008), which also applies under turbulent flows.

Similarly, "anemotaxis" is defined as "orientation or movement in response to a fluid flow", typically air or water.

Many searching strategies based on chemotaxis and anemotaxis have been inspired by bacteria or animal behaviour, which use olfaction for foraging or mating, for instance: *Eschericia coli* bacteria (Russell et al. 2003), lobsters (Grasso et al. 2000), blue crabs (Webster, Volyanskyy & Weissburg 2012), ants (Russell 1999), moths (Pyk et al. 2006; Edwards et al. 2005), dung beetles (Ishida et al. 1994), among others (Murlis, Elkinton & Carde 1992). These approaches are known as bio-inspired algorithms.

1.5.3.1 Reactive plume tracking strategies

All bio-inspired algorithms can be classified in the group of reactive plume tracking strategies. Basically, in these strategies, the gas source localization problem is divided in three stages: gas finding, plume tracking and source declaration (Lilienthal, Loutfi & Duckett 2006; Kowadlo & Russell 2008).

In the gas finding stage, the robot tries to make contact with the odour; in the plume tracking stage, the robot moves reactively along or within the plume until it finds the source, which is decided in the source declaration stage and thus the search is finished.

Four reactive chemotaxis algorithms are implemented and evaluated in (Russell et al. 2003). Bio-inspired approaches based on the behaviour of *Eschericia coli* bacteria, silkworm moth *Bombyx mori* and dung bettle, and a gradient-based approach based on the Braitenberg vehicle (Braitenberg 1986) are compared in the paper.

Three algorithms are adapted from the behaviour of *E. coli* bacteria and cooperation between robots is added in (Lytridis, Virk & Kadar 2005). It is shown that using multiple cooperative robots is more efficient to localize the source than using a single robot.

In (Ishida et al. 2004), the first autonomous robotic odour localization in 3D using a blimp is reported. An array of ten tin-oxide sensors is mounted over the balloon surface. It is seen that the sensors facing toward the gas source exhibit "active" responses whereas the sensors behind the large balloon body show "quiet" responses. Moreover, two algorithms are implemented and tested experimentally.

In order to improve the localization of the source, anemotaxis is used together with chemotaxis. Not all strategies use anemometric sensors, but if the fluid flow can be indirectly sensed (distributed chemical sensors), it is considered that the algorithm implements anemotaxis.

The zigzag approach based on the dung beetle behaviour was implemented above ground (Ishida et al. 1994; Ishida et al. 1996) and underwater (Farrell, Pang & Li 2005).

The work by Farrell et al. appears to be the largest scale odour localization experiment (hundreds of square metres) by at least two orders of magnitude greater than typical experiments.

A plume-centered upwind search strategy was implemented successfully by (Russell et al. 1995); later Ishida et al. (Ishida et al. 2005) integrated a similar strategy into a multiphase approach which improved the zigzag method.

Inspired by the mating dance of the silkworm moth *Bombix mori*, odour compasses were implemented for two-dimensional (Nakamoto, Ishida & Moriizumi 1996) and three-dimensional localization (Ishida et al. 1999). In these works, the fluid flow is actively manipulated. Similar to the animal behaviour of moving its wings to draw air samples with pheromones from the front to the antennae, multiple sensors were placed in front of a propeller fan in order to improve sensitivity.

A search strategy based on the surge-cast-turn behaviour of a silkworm moth using one chemical sensor and one anemometer was implemented by (Marques, Nunes & Almeida 2002). In (Russell et al. 2003) and (Pyk et al. 2006) two chemical sensors are used. The results showed the method to be more effective than the gradient-based and *E. coli* methods. The silkworm strategy was adapted to be used in a swarm of robots (Hayes, Martinoli & Goodman 2002). The method successfully located the odour source in simulations and in practical experiments. Furthermore, similar to other works (Marjovi & Marques 2011), it is seen that chemical sources are localized faster and more robustly when using multiple mobile robots.

A reactive algorithm for a swarm implementing a silkworm moth casting to find the plume, called "fluxotaxis", was presented by (Zarzhitsky, Spears & Spears 2005). Information about fluid velocity, chemical concentration and mass flux is used in algorithm to find the source and not a local maximum in the averaged chemical distribution. Successful simulation results are presented in the article; nevertheless practical experiments are required so as to establish fluxotaxis as an effective method.

An alternative chemical reactive method, called "infotaxis", is presented in (Vergassola, Villermaux & Shraiman 2007). In this approach information plays a role similar to concentration in chemotaxis. Odour patches are expected to be found intermittently in the medium, and then information is sparse. It is assumed that more odour patches are found close to the source. In this strategy, the algorithm chooses the move that locally maximizes the expected rate of information gain. The rate of acquisition of information is quantified by the rate of reduction of entropy, which depends on the probability for the location of the source. This probability is updated as odour patches are encountered. The more information, the less the expected time to localize the source. Results from simulations and practical experiments shown that

infotactic trajectories feature "zigzagging" and "casting" paths similar to those observed in the flight of moths.

Four different searching strategies were implemented in turbulent flows using swarms of robots and tested through simulations in (Lo Iacono 2010). Two strategies only consider the rate of odour patches and the other two also consider the variation of odour concentrations. It is suggested that the algorithms which use concentration information are more effective to localize the source.

As it is pointed in previously cited works and in (Mcgill & Taylor 2011), swarming algorithms have provided more robust solutions to the single-source localization problem; moreover, they appear to be appropriate methods for the multiple-source problem. However, little attention has been paid to it. McGill and Taylor provides a review of methods focusing on the detection of multiple sources (Mcgill & Taylor 2011).

In (Cui et al. 2004) the Biasing Expansion Swarm Approach (BESA) is presented in order to localize an unspecified number of hazardous chemical sources in a large, unknown area. Inspired in the social foraging behaviour of *E. coli* bacteria, a robust swarming algorithm in the presence of multiple nutrient sources is presented in (Liu & Passino 2002). Also inspired by the chemotaxis of *E. coli* bacteria, the Biased Random Walk (BRW) algorithm applied on a large robotic network to localize multiple gradient sources is described in (Dhariwal, Sukhatme & Requicha 2004). Another bio-inspired algorithm is the Glowworm Swarm Optimization (GSO), which features an adaptive decision domain which enables the formation of subgroups in the swarm where the goal is to partition the swarm to track multiple sources concurrently (Krishnanand & Ghose 2009).

One of the main problems in reactive plume tracking strategies is that a source declaration stage should be implemented so as to decide when finishing the search. In general, in order to avoid too many complications in the published experiments, the tracking is stopped simply when the robots came near to the chemical sources. Concerning the source declaration problem, very little work can be found in the literature. Some works try to find indicators to decide when a gas source has been found, for instance analyzing series of concentration measurements (Lilienthal et al. 2004; Lilienthal et al. 2006; Loutfi & Coradeschi 2006). In other works additional capabilities like vision are added to the robotic systems; using vision it is possible to recognize obstacles and determine, from the sensor readings, if it is a candidate to gas source or not (Martinez & Perrinet 2002; Kowadlo et al. 2006; Ishida et al. 2006).

1.5.3.2 Plume modelling

In some applications it is sufficient to know the distribution of a target chemical in a given environment without tracking the entire length of the plume. Unlike reactive plume tracking approaches, one interesting characteristic of plume modelling methods is that it is not strictly necessary to track the plume to localize the chemical source. Usually the source location is a parameter which is integrated within the model and is updated in real-time. This implies that the robots could be intended for other tasks in addition of finding the source (Larionova et al. 2006). Therefore, the problem of source declaration is avoided since the end of the search is not determined by any tracking algorithm.

Apart from the information from the sensors, typically chemical and anemometric, these approaches integrate additional knowledge about the plume dispersion. Since it is not feasible to model the turbulent wind and gas distribution in a realistic environment with the current technology, plume information can be modelled in a computationally inexpensive way using analytic gas distribution models (Ishida, Nakamoto & Moriizumi 1998; Marques, Nunes & Almeida 2002) or stochastic/probabilistic methods (Pang & Farrell 2006). As in the case of vision, adding additional knowledge to the robotic system helps to solve the source localization problem.

The first example of plume modelling was reported in (Ishida, Nakamoto & Moriizumi 1998). The algorithm was designed to remotely localize the odour source. The model parameters, including the source location, are continuously updated while the robot zigzags toward the source.

An analytic gas distribution model is integrated into plume tracking in (Marques, Nunes & Almeida 2002). As the robot navigates, the model is continuously updated. Although in these two works the robot tracks the plume, it is important to note that the localization of the source is decoupled of the algorithm used to guide the robot. In fact, an algorithm to move the robot would not be strictly needed.

This is the idea followed in (Lilienthal & Duckett 2004), where predefined paths are implemented in a robot in order to map an area with a gas chemical source. Gaussian weighting functions are used to model the decreasing likelihood that a particular reading represents the true concentration with respect to the distance from the point of measurement. A map of the area is obtained and updated in real-time. The proposed method was later improved so that not only the mean gas distribution was estimated but also the predictive variance (Lilienthal et al. 2009). It is observed that the variance map provides more accurate information about the source location than the mean concentration map. In (Stachniss, Plagemann & Lilienthal 2009) two-

dimensional gas distribution models are learnt from the data using sparse Gaussian process mixtures. This method allows to model areas of high concentration and areas with smooth background signal, thus improving the accuracy of the gas concentration prediction. In (Reggente & Lilienthal 2010) is shown that airflow measurements are useful to improve the gas distribution mapping. Also in this work multiple gas sensors are placed at different heights so as to perform a three-dimensional mapping. Alternatively, three-dimensional gas mapping is done in (Ishida 2009) using a blimp. The flying robot moved randomly and collected gas concentration measurements at different locations in the environment.

In (Reggente et al. 2010) is proposed to use a gas distribution mapping algorithm as an effective method to measure air-pollutant distribution in a city.

Instead of using analytic gas distribution models, plume modelling using stochastic methods has also been proposed in the literature. These techniques show a great potential and bring probability to the field of odour localization.

Farrell published an approach based on computationally efficient Hidden Markov Models (HMM) (Farrell 2003). The main idea behind the algorithm consists in implementing a "Hidden Markov Plume Model" and estimating the likelihood of odour detection, the likelihood of source location, the most likely path taken by an odour to a given location and the path between two points most likely to result in odour detection. It uses the measurements from the chemical sensors and anemometers.

The algorithm was later improved to incorporate Bayesian inference and to be more efficient (Pang & Farrell 2006). This new approach takes into account the sequence of detection/non-detection events and fluid flow measurements along the trajectory of the robot to build a source probability map. From the probability map, which is updated recursively using Bayesian inference, the most likely source location is estimated in real-time. Dispersion of chemical filaments is modelled as a random walk (due to velocity fluctuations) superimposed on downwind advection (due to mean velocity). Experiments using real-world data show that this approach can be effective to predict the likely source location of a chemical source over a large area, with increasing accuracy, as more measurements are gathered. In addition, the source probability map provided by Bayesian reasoning can be used to design new tracking strategies in order to minimize the searching time.

In (Li et al. 2011) a particle filter algorithm is proposed. In this work the plume is modelled as a collection of particles, each one with its own location and with a certain weight. The weights are updated while the robot is exploring the area and the location of the source is estimated in real-time. Experimental results show that the proposed

approach performs better than the Bayesian-inference-based algorithm by Pang and Farrell (Pang & Farrell 2006).

Single-source tracking has also been applied to swarms of robots or mobile sensor networks using probabilistic models, for instance Bayesian-based distributed target tracking (Zou & Chakrabarty 2007) or integrating a network of cooperative Kalman filters with a flocking-based mobility model (Olfati-Saber 2007).

Other proposals are based on Particle Swarm Optimization (PSO) (Kennedy & Eberhart 1995; Jatmiko, Sekiyama & Fukuda 2007; Jatmiko et al. 2008), where swarms of particles, or robots, move in a virtual space to find an optimal solution. It is assumed that robots can communicate with each other and evaluate their fitness to the solution at each location. Collisions are avoided by treating other robots or obstacles as repulsive forces.

Modifications have been performed in order to adapt the algorithm to the multi-source problem (Marques, Nunes & Almeida 2006). Additionally, Jakuba and Ferri et al. have adapted Bayesian occupancy grid mapping to localize multiple gas sources (Jakuba 2007; Ferri et al. 2011). They improved the work by Pang and Farrell (Pang & Farrell 2006) by eliminating the assumption of a single-source and modifying the updating rule applied to the source probability map so as to account for the low prior probabilities in multi-source searches.

1.5.4 Turbulence dominated weak fluid flow

In certain indoor environments like narrow corridors, factory buildings, tunnels, caves or mines, strong and constant airflows are not likely to found. In these cases, difficulties associated with variations in temperature, humidity and flow arise. The fluid flow can be very small, sometimes smaller than the detection limit of ultrasonic anemometers; moreover, due to the walls surrounding the environment, circulating convective airflows are generated and if the temperature distribution is complicated, the flow field becomes highly irregular (Trincavelli et al. 2008).

Since it appears to be too difficult to develop a chemical source localization algorithm which could be used in any scenario, specialized robots designed for a specific target application are often used in enclosed environments, for instance using one robot (Russell 2001) or multiple robots (Marjovi & Marques 2011).

In (Kowadlo & Russell 2006) naive physics are used to model the airflow in the environment. This provides approximate solutions with enough degree of accuracy for the source localization task. Using the modelled airflow map, the robot moves to

specific target locations and the likelihood that the odour was emitted from these locations is computed.

In (Lilienthal et al. 2001) a robot to locate an odour source without the presence of a strong airflow is presented. Experiments were performed in two standard corridors. The results showed that the location of the source corresponds only roughly to the location of the maximal instantaneous measured concentration. Moreover, the averaged concentration peak does not remain stationary at the location of the source due to convective airflow currents. A version of the algorithm for building gas concentration gridmaps (Lilienthal & Duckett 2004) but focused on weak fluid flows is presented in (Lilienthal, Streichert & Zell 2005). In this work the robot movement is not controlled by a reactive tracking strategy and a Gaussian gas distribution model is used to build a map to localize the source.

An adaptation of the silkworm moth algorithm, for weak fluid flows, avoiding the surge behaviour, which requires knowledge of the airflow direction, is described in (Lilienthal, Reimann & Zell 2003). Purnamadjaja and Russell implemented and adapted *E. coli* bacteria algorithm in a swarm of cooperative robots for pushing a dead robot (Purnamadjaja & Russell 2005). Experimental results show that the method is able to localize the source effectively and reliably, from a relatively large distance, with only one chemical sensor, with no airflow information and without the presence of a strong airflow.

1.5.5 Gas identification and localization in a real environment

Although previously cited algorithms focus on the localization of a gas chemical source or multiple sources, using one robot or multiple robots, the majority of them only consider the detection of a specific target substance. However, in a real environment, where multiple substances are expected to be present, the sensors (unless selective sensors are used) will respond to many different chemical compounds, regarded as interfering, apart from the substance of interest.

In many works, when using one sensor, the typical procedure consists in setting a threshold level to the sensor signal in order to minimize the impact of interfering compounds in the localization of the source. If the sensor reading is above the threshold, it is considered that the robot is making contact with the plume.

Deciding the threshold level is another issue which has not been addressed properly in the literature. This issue is especially important when the gas source that has to be detected is weak compared to the background of interfering substances. Setting the threshold too high will cause too many miss-detections, setting the threshold too low will cause too many false alarms; thus leading the algorithm to fail.

Most of the robotic systems which incorporate more than one sensor use gas sensor arrays. Typically, each particular substance is associated with a certain pattern of response in the sensor array. Therefore, when using gas sensor arrays, it is possible to discriminate among different substances and separate the contribution of the target from the contribution of the interfering substances.

However, not many works have addressed this issue. The most used kind of gas sensors is metal-oxide sensors because of their reasonably high sensitivities, good response times (typically less than five seconds), low price and long lifetimes. Nevertheless, the recovery time of these sensors is very slow (typically tens of seconds), which is not enough to capture the instantaneous plume fluctuations encountered in turbulent flows (typically less than one second), that can reach a bandwidth of 100 Hz.

In order to overcome this problem, instead of setting a threshold to the direct response, the transient signal of the sensors, which carries important information about instantaneous plume characteristics, is used in (Ishida et al. 2005); however, the discrimination problem among different substances is not addressed.

As far as we know, only (Marques, Nunes & Almeida 2002) used transient signals of a gas sensor array for tracking a target source (ethanol) in the presence of an interfering source (methanol).

Additionally, some recent works from the group of Prof. Dr. Achim J. Lilienthal have been published focusing on the identification and localization of odours in a real environment. In these works, instead of tracking the plume, a plume mapping is performed. For instance, in (Loutfi et al. 2009) the transient response of four MOX sensors is used to classify odours (acetone, ethanol and air) and then creating a combined map where the multiple trained substances can be identified. Experimental results using one robot in indoors and outdoors are presented.

In the thesis published by Dr. Marco Trincavelli (Trincavelli 2010) the problem of gas discrimination in a real environment using a mobile robot is also addressed. Experimental results in indoors, outdoors, using predefined paths or different searching strategies and with the presence of different gas sources (ethanol, acetone, isopropyl, 2-propanol and air) are presented. MOX sensors are also used in the thesis and what features are better for odour discrimination is discussed in detail.

Even though, in these works, different substances are discriminated and a gas distribution mapping separating the contributions of the different substances is

performed, there is not any quantitative step, thus only qualitative results are provided.

The use of MOX sensors for real-time applications limits the capabilities of the robotic system. Since the recovery time is slow, instantaneous concentrations can be hidden after a peak of high concentration, thus losing important information for quantification. This problem is solved in (Hernández-Bennetts, Lilienthal & Trincavelli 2012) using six MOX sensors and the readings of a PID sensor to obtain calibrated measurements. First, the MOX readings at a certain location are classified to come from one of the trained substances (ethanol, 2-propanol and fresh air); secondly, the calibrated PID readings are used to create calibrated maps of each one of the substances. Therefore, concentration information from the PID is fused with the MOX readings within the gas mapping algorithm.

Thereby, as it has also been pointed before (Lo Iacono 2010), instead of having binary detections above a certain threshold, integrating concentration information within the source localization algorithm could be useful to improve the localization of the gas source, especially under the presence of interfering substances. However, to the best of our knowledge, apart from the work by (Hernández-Bennetts, Lilienthal & Trincavelli 2012), there are no more proposals in the literature.

1.6 Summary

This chapter introduces the general problems of gas detection, identification and source localization in different scenarios. In some of these scenarios the sample can be taken and analyzed, but in many others an exploration of the area must be performed because the localization of the gas source is unknown. This exploration can be carried out using multiple sensors (fixed or mobile).

Different sensing technologies have been successfully used to detect and identify different chemical substances, such as explosives or volatile organic compounds (VOCs). Among these technologies, mobility based analyzers provide fast responses with high sensitivity.

However, IMS instruments are not exempt of problems. Typically, they provide moderate selectivity, appearing overlapped peaks in the spectra. Moreover, the presence of humidity makes peaks wider, thus worsening the resolving power and the resolution. Furthermore, the response of IMS is non-linear as substance concentration increases and more than one peak can appear in the spectra due to the same compound.

In order to deal with these issues, IMS spectra can be processed using multivariate techniques. The more relevant approaches to analyze the spectra qualitatively and

quantitatively have been reviewed. Moreover, techniques to classify chemical samples have been presented.

Despite the advantages of using multivariate signal and data processing in IMS, the application of multivariate signal and data processing to IMS spectra is limited. Understanding the advantages and limitations of chemometrics for IMS still requires further investigations.

In the last part of this introduction, the problem of gas dispersion in a real environment is presented. Different strategies have been proposed in the literature adapted to different fluid flows so as to solve the source localization problem. It is shown that the use of multiple mobile sensors help to localize the gas source.

Among the different algorithms, plume modelling approaches provide strategies to localize the source without performing plume tracking, thus the mobile robots can be used for other tasks.

In general, the algorithms only focused on one target substance and only its presence or absence is important. Thereby, the source localization is usually based on a binary threshold.

With this regard, little work has been done in the discrimination of multiple substances in a real environment and in the use of the concentration measurements provided by the sensors, thus not discarding useful information.

1.7 References

Abbaspour, M & Mansouri, N 2005, 'City hazardous gas monitoring network', *Journal of Loss Prevention in the Process Industries*, vol. 18, no. 4-6, pp. 481-487.

ACM, *http://www.appliedsensor.com/news/2012/AppliedSensor_06062012.html*, AppliedSensor.

Ali, EMA, Edwards, HGM & Scowen, IJ 2009, 'Raman spectroscopy and security applications: the detection of explosives and precursors on clothing', *Journal of Raman Spectroscopy*, vol. 40, no. 12, pp. 2009-2014.

Alonso, M, Santos, JP, Hontanon, E & Ramiro, E 2009, 'First Differential Mobility Analysis (DMA) Measurements of Air Ions Produced by Radioactive Source and Corona', *Aerosol and Air Quality Research*, vol. 9, no. 4, pp. 453-457.

Amari, SI, Cichocki, A & Yang, H 1996, 'A new learning algorithm for blind source separation', *Advances in Neural Information Processing Systems*, vol. 8, pp. 757-763.

Antunes, MC, Simao, JE, Duarte, AC, Esteban, M & Tauler, R 2002, 'Application of multivariate curve resolution to the voltammetric study of the complexation of fulvic acids with cadmium(II) ion', *Anal. Chim. Acta*, vol. 459, no. 2, pp. 291-304.

Armenta, S, Alcala, M & Blanco, M 2011, 'A review of recent, unconventional applications of ion mobility spectrometry (IMS)', *Analytica Chimica Acta*, vol. 703, no. 2, pp. 114-123.

Armenta, S & Blanco, M 2011, 'Pros and cons of benzodiazepines screening in human saliva by ion mobility spectrometry', *Analytical and Bioanalytical Chemistry*, vol. 401, no. 6, pp. 1935-1948.

Armenta, S & Blanco, M 2012, 'Ion Mobility Spectrometry: A Comprehensive and Versatile Tool for Occupational Pharmaceutical Exposure Assessment', *Analytical Chemistry*, vol. 84, no. 10, pp. 4560-4568.

Azzouz, T & Tauler, R 2008, 'Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples', *Talanta*, vol. 74, no. 5, pp. 1201-1210.

Babis, JS, Sperline, RP, Knight, AK, Jones, DA, Gresham, CA & Denton, MB 2009, 'Performance evaluation of a miniature ion mobility spectrometer drift cell for application in hand-held explosives detection ion mobility spectrometers', *Analytical and Bioanalytical Chemistry*, vol. 395, pp. 411-419.

Back, D & Weigend, AS 1997, 'A first application of independent component analysis to extracting structure from stock returns', *International Journal on Neural Systems*, vol. 8, no. 4, pp. 473-484.

Baffi, G, Martin, EB & Morris, AJ 1999, 'Non-linear projection to latent structures revisited: the quadratic PLS algorithm', *Computers & Chemical Engineering*, vol. 23, no. 3, pp. 395-411.

Bakkum, EA & Duijm, NJ 2005, 'Vapour cloud dispersion', in *Methods for the calculation of physical effects 'Yellow Book'*, 3rd edn, The Hague, pp. 4.1-4.139.

Barclay, VJ, Bonner, RF & Hamilton, IP 1997, 'Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression', *Analytical Chemistry*, vol. 69, no. 1, pp. 78-90.

Barker, M & Rayens, W 2003, 'Partial least squares for discrimination', *Journal of Chemometrics*, vol. 17, no. 3, pp. 166-173.

Baumbach, J 2006, 'Process analysis using ion mobility spectrometry', *Analytical and Bioanalytical Chemistry*, vol. 384, no. 5, pp. 1059-1070.

Baumbach, JI, Sielemann, S, Xie, ZY & 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*, vol. 75, no. 6, pp. 1483-1490.

Bell & Sejnowski, T 1995, 'An information-maximization approach to blind separation and blind deconvolution', *Neural Computation*, vol. 7, pp. 1129-1159.

Bell, SE, Ewing, RG, Eiceman, GA & Karpas, Z 1994, 'Atmospheric pressure chemical ionization of alkanes, alkenes, and cycloalkanes', *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 3, pp. 177-185.

Bell, WJ & Tobin, TR 1982, 'Chemo-Orientation', *Biological Reviews*, vol. 57, no. 2, pp. 219-260.

Berkowitz, Z, Haugh, GS, Orr, MF & Kaye, WE 2002, 'Releases of hazardous substances in schools: Data from the Hazardous Substances Emergency Events Surveillance system, 1993-1998', *Journal of Environmental Health*, vol. 65, no. 2, pp. 20-27.

Bishop, CM 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc.

Borah, DK & Balagopal, A 2004, 'Localization and tracking of multiple near-field sources using randomly distributed sensors', in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, pp. 1323-1327 Vol.2.

Borsdorf, H, Mayer, T, Zarejousheghani, M & Eiceman, GA 2011, 'Recent Developments in Ion Mobility Spectrometry', *Applied Spectroscopy reviews*, vol. 46, no. 6, pp. 472-521.

Borsdorf, H, Stone, JA & Eiceman, GA 2005, 'Gas phase studies on terpenes by ion mobility spectrometry using different atmospheric pressure chemical ionization techniques', *International Journal of Mass Spectrometry*, vol. 246, no. 1-3, pp. 19-28.

Bourgeois, W, Romain, A-C, Nicolas, J & Stuetz, RM 2003, 'The use of sensor arrays for environmental monitoring: interests and limitations', *Journal of Environmental Monitoring*, vol. 5, no. 6, pp. 852-860.

Braitenberg, V 1986, *Vehicles: Experiments in synthetic psychology*, MIT press.

Buffo, RA & Cardelli-Freire, C 2004, 'Coffee flavour: an overview', *Flavour and Fragrance Journal*, vol. 19, no. 2, pp. 99-104.

Buscemi, L & Prati, M 1994, 'Cellular Robotics: Behaviour in Polluted Environments', in *Proceedings of the 2nd International Symposium on Distributed Autonomous Robotic Systems*, Saitama.

Buxton, TL & Harrington, PD 2003, 'Trace explosive detection in aqueous samples by solid-phase extraction ion mobility spectrometry (SPE-IMS)', *Applied Spectroscopy*, vol. 57, no. 2, pp. 223-232.

Buxton, TL & Harrington, PdB 2001, 'Rapid multivariate curve resolution applied to identification of explosives by ion mobility spectrometry', *Analytica Chimica Acta*, vol. 434, pp. 269-282.

Calvo, D, Tort, N, Salvador, JP, Marco, MP, Centi, F & Marco, S 2011, 'Preliminary study for simultaneous detection and quantification of androgenic anabolic steroids using ELISA and pattern recognition techniques', *Analyst*, vol. 136, no. 19, pp. 4045-4052.

Cao, L & Harrington, PdB 2004, 'ALS to modeling of ion mobility spectra', *International Journal for Ion Mobility Spectrometry*, vol. 7, no. 1, pp. 9-11.

Cao, L, Harrington, PdB & Liu, J 2005, 'SIMPLISMA and ALS Applied to Two-Way Nonlinear Wavelet Compressed Ion Mobility Spectra of Chemical Warfare Agent Simulants', *Analytical Chemistry*, vol. 7, pp. 2575-2586.

Cichoki, A, Zdunek, R, Phan, AH & Amari, S (eds) 2009, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley & Sons, Chichester.

Coombes, KR, Tsavachidis, S, Morris, JS, Baggerly, KA, Hung, MC & Kuerer, HM 2005, 'Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform', *Proteomics*, vol. 5, no. 16, pp. 4107-4117.

Crimaldi, JP, Wiley, MB & Koseff, JR 2002, 'The relationship between mean and instantaneous structure in turbulent passive scalar plumes', *Journal of Turbulences*, vol. 3, no. 14, pp. 1-24.

Cui, X, Hardin, CT, Ragade, RK & Elmaghraby, AS 2004, 'A swarm approach for emission sources localization', in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pp. 424-430.

Cyranose320, *http://www.intopsys.com/products/cyranose.html?gclid=CM7h_b7BwrMCFUbKtAodn QUA0Q*, Intelligent Optical Systems.

ChemPro100, *http://www.environics.fi/index.php/chemical-detection/chempro100i*, Environics.

Chen, G 2003, Real-Time Wavelet Compression and Self-Modeling Curve Resolution for Ion Mobility Spectrometry, thesis, Ohio.

Chen, P 2008, Applications of Chemometric Algorithms to Ion Mobility Spectrometry and Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry, thesis, Ohio University.

Chen, P & Harrington, PB 2008, 'Discriminant Analysis of Fused Positive and Negative Ion Mobility Spectra Using Multivariate Self-Modeling Mixture Analysis and Neural Networks', *Applied Spectroscopy*, vol. 62, no. 2, pp. 133-141.

Daum, KA, Watrous, MG, Neptune, MD, Michael, DI, Hull, KJ & Evans, JD 2006, *Data for First Responder Use of Photoionization Detectors for Vapor Chemical Constituents*, INL/EXT-05-00165; TRN: US200801%%351.

De Vito, S & Fattoruso, G 2012, 'Wireless Chemical Sensor Networks for Air Quality Monitoring', in *The 14th International Meeting on Chemical Sensors (IMCS 2012)*, Nuremberg, Germany, pp. 641 - 644.

De Vito, S, Fattoruso, G, Liguoro, R, Oliviero, A, Massera, E, Sansone, C, Casola, V & Di Francia, G 2011, 'Cooperative 3D Air Quality Assessment with Wireless Chemical Sensing Networks', *Procedia Engineering*, vol. 25, no. 0, pp. 84-87.

Dhariwal, A, Sukhatme, GS & Requicha, AAG 2004, 'Bacterium-inspired robots for environmental monitoring', in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, pp. 1436-1443 Vol.2.

Disaster_Information_Management_Research_Center, *http://sis.nlm.nih.gov/enviro/chemicalwarfare.html#a1*, U.S Department of Health and Human Services.

Distante, C, Indiveri, G & Reina, G 2009, 'An application of mobile robotics for olfactory monitoring of hazardous industrial sites', *Industrial Robot-an International Journal*, vol. 36, no. 1, pp. 51-59.

Donoho, DL 1995, 'De-Noising by Soft-Thresholding', *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613-627.

Dworzanski, JP, Kim, MG, Snyder, AP, Arnold, NS & Meuzelaar, HLC 1994, 'Performance Advances in Ion Mobility Spectrometry through combination with high-speed vapor

sampling, preconcentration and separation techniques', *Analytica Chimica Acta*, vol. 293, no. 3, pp. 219-235.

Edwards, S, Rutkowski, AJ, Quinn, RD & Willis, MA 2005, 'Moth-Inspired Plume Tracking Strategies In Three-Dimensions', in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pp. 1669-1674.

Eiceman, GA 2002, 'Ion-mobility spectrometry as a fast monitor of chemical composition', *TrAC Trends in Analytical Chemistry*, vol. 21, no. 4, pp. 259-275.

Eiceman, GA 2002, 'Ion mobility spectrometry as detector and sensor for chemical warfare agents and toxic industrial chemicals', *Abstracts of Papers of The American Chemical Society*, vol. 224, no. 1, p. U145.

Eiceman, GA & Karpas, Z 2005, *Ion Mobility Spectrometry*, second edn, CRC Taylor and Francis Group, Boca Raton.

Eiceman, GA, Krylov, EV, Krylova, NS, Nazarov, EG & Miller, RA 2004, 'Separation of Ions from Explosives in Differential Mobility Spectrometry by Vapor-Modified Drift Gas', *Analytical Chemistry*, vol. 76, no. 17, pp. 4937-4944.

Eiceman, GA, Sowa, S, Lin, S & Bell, SE 1995, 'Ion mobility spectrometry for continuous on-site monitoring of nicotine vapors in air during the manufacture of transdermal systems', *Journal of Hazardous Materials*, vol. 43, no. 1-2, pp. 13-30.

EntryScan4, *http://www.morpho.com/IMG/pdf/MDI_EntryScan_4_DAT.pdf*, Morpho Detection.

Espy, M, Flynn, M, Gomez, J, Hanson, C, Kraus, R, Magnelind, P, Maskaly, K, Matlashov, A, Newman, S, Owens, T, Peters, M, Sandin, H, Savukov, I, Schultz, L, Urbaitis, A, Volegov, P & Zotev, V 2010, 'Ultra-low-field MRI for the detection of liquid explosives', *Superconductor Science and Technology*, vol. 23, no. 3, p. 034023.

Esteban, M, Ariño, C, Díaz-Cruz, JM, Díaz-Cruz, MS & Tauler, R 2000, 'Multivariate curve resolution with alternating least squares optimisation: a soft-modelling approach to metal complexation studies by voltammetric techniques', *TrAC Trends in Analytical Chemistry*, vol. 19, no. 1, pp. 49-61.

Ewing, RG, Atkinson, DA, Eiceman, GA & Ewing, GJ 2001, 'A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds', *Talanta*, vol. 54, pp. 515-529.

Ewing, RG, Eiceman, GA & Stone, JA 1999, 'Proton-bound cluster ions in ion mobility spectrometry', *International Journal of Mass Spectrometry*, vol. 193, no. 1, pp. 57-68.

Ewing, RG & Miller, CJ 2001, 'Detection of Volatile Vapors Emitted from Explosives with a Handheld Ion Mobility Spectrometer', *Field Analytical Chemistry and technology*, vol. 5, no. 5, pp. 215-221.

Fackrell, JE & Robins, AG 1982, 'Concentration fluctuations and fluxes in plumes from point sources in a turbulent boundary layer', *Journal of Fluid Mechanics*, vol. 117, pp. 1-26.

Farrell, JA 2003, 'Plume Mapping via hidden markov methods', *IEEE Transactions on systems, man, and cybernetics*, vol. 33, no. 6, pp. 850-863.

Farrell, JA, Murlis, J, Long, X, Li, W & Cardé, RT 2002, 'Filament-based atmospheric dispersion model to achieve short time-scale structure of odor plumes', *Environmental Fluid Mechanics*, vol. 2, pp. 143-169.

Farrell, JA, Pang, S & Li, W 2005, 'Chemical Plume Tracing via an Autonomous Underwater Vehicle', *IEEE Journal of Oceanic Engineering*, vol. 30, no. 2, pp. 428-442.

Fernandez-Lima, F, Kaplan, D, Suetering, J & Park, M 2011, 'Gas-phase separation using a trapped ion mobility spectrometer', *International Journal for Ion Mobility Spectrometry*, vol. 14, no. 2-3, pp. 93-98.

Fernandez-Lima, FA, Kaplan, DA & Park, MA 2011, 'Note: Integration of trapped ion mobility spectrometry with mass spectrometry', *Review of Scientific Instruments*, vol. 82, no. 12, pp. 126106-126106-3.

Ferri, G, Jakuba, MV, Mondini, A, Mattoli, V, Mazzolai, B, Yoerger, DR & Dario, P 2011, 'Mapping multiple gas/odor sources in an uncontrolled indoor environment using a Bayesian occupancy grid mapping based method', *Robotics and Autonomous Systems*, vol. 59, no. 11, pp. 988-1000.

Fiorelli, E, Bhatta, P & Leonard, NE 2003, 'Adaptive sampling using feedback control of an autonomous underwater glider fleet ', in *Proc. 13th Int. Symposium on Unmanned Untethered Submersible Tech*.

Fisher, RA 1936, 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, vol. 7, pp. 179-188.

FOX4000, *http://www.alpha-mos.com/products_technology/electronic_nose.php*, Alpha MOS.

Fraga, CG, Kerr, DR & Atkinson, DA 2009, 'Improved quantitative analysis of ion mobility spectrometry by chemometric multivariate calibration', *Analyst*, vol. 134, no. 11, pp. 2329-2337.

Furton, KG & Myers, LJ 2001, 'The scientific foundation and efficacy of the use of canines as chemical detectors for explosives', *Talanta*, vol. 54, no. 3, pp. 487-500.

Gan, F, Ruan, G & Mo, J 2006, 'Baseline correction by improved iterative polynomial fitting with automatic threshold', *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 59-65.

Gardner, JW & Bartlett, PN 1994, 'A brief history of electronic noses', *Sensors and Actuators B: Chemical*, vol. 18, no. 1â€"3, pp. 210-211.

Gardner, JW, Shin, HW & Hines, EL 2000, 'An electronic nose system to diagnose illness', *Sensors and Actuators B: Chemical*, vol. 70, no. 1, pp. 19-24.

Garrido-Delgado, R, Arce, L, Guamán, AV, Pardo, A, Marco, S & Valcárcel, M 2011, 'Direct coupling of a gas-liquid separator to an ion mobility spectrometer for the classification of different white wines using chemometrics tools', *Talanta*, vol. 84, no. 2, pp. 471-479.

GC-IMS, *http://www.gas-dortmund.de/Products/GC-IMS/1_367.html*, G.A.S.

GDA2, *http://www.airsense.com/en/products/gda-2/*, Airsense Analytics.

Gemperline, PJ & Cash, E 2003, 'Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions', *Analytical Chemistry*, vol. 75, no. 16, pp. 4236-4243.

Genovese, V, Dario, P, Magni, R & Odetti, L 1992, 'Self Organizing Behavior And Swarm Intelligence In A Pack Of Mobile Miniature Robots In Search Of Pollutants', in *Intelligent Robots and Systems, 1992., Proceedings of the 1992 IEEE/RSJ International Conference on*, pp. 1575-1582.

Giles, K, Williams, JP & Campuzano, I 2011, 'Enhancements in travelling wave ion mobility resolution', *Rapid Communications in Mass Spectrometry*, vol. 25, no. 11, pp. 1559-1566.

Golub, G & Reinsch, C 1970, 'Singular value decomposition and least squares solutions', *Numerische Mathematik*, vol. 14, no. 5, pp. 403-420.

Grasso, FW, Consi, TR, Mountain, DC & Atema, J 2000, 'Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges', *Robotics and Autonomous Systems*, vol. 30, no. 1-2, pp. 115-131.

Gutierrez-Osuna, R 2002, 'Pattern Analysis for Machine Olfaction: A Review', *Ieee Sensors Journal*, vol. 2, no. 3, pp. 189-202.

Hargreaves, MD, Page, K, Munshi, T, Tomsett, R, Lynch, G & Edwards, HGM 2008, 'Analysis of seized drugs using portable Raman spectroscopy in an airport environment—a proof of principle study', *Journal of Raman Spectroscopy*, vol. 39, no. 7, pp. 873-880.

Hastie, T, Tibshirani, R & Friedman, JH 2003, *The Elements of Statistical Learning*, Springer.

Hayes, AT, Martinoli, A & Goodman, RM 2002, 'Distributed odor source localization', *Sensors Journal, IEEE*, vol. 2, no. 3, pp. 260-271.

Hernández-Bennetts, V, Lilienthal, A & Trincavelli, M 2012, 'Creating true gas concentration maps in presence of multiple heterogeneous gas sources', in *Proceedings of the IEEE Sensors Conference* Taipei, Taiwan, p. to appear.

Hernández-Bennetts, V, Lilienthal, AJ, Khaliq, AA, Pomareda, V & Trincavelli, M 2012, 'Gasbot: A mobile robotic platform for methane leak detection and emission monitoring ', in *IROS Workshop on Robotics for Environmental Monitoring*, Vilamoura, Algarve, Portugal, p. to appear.

Hilton, CK, Krueger, CA, Midey, AJ, Osgood, M, Wu, J & Wu, C 2010, 'Improved analysis of explosives samples with electrospray ionization-high resolution ion mobility spectrometry (ESI-HRIMS)', *International Journal of Mass Spectrometry*, vol. 298, pp. 64-71.

Hill, HH, Siems, WF, Stlouis, RH & McMinn, DG 1990, 'Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 62, no. 23, pp. A1201-A1209.

Hill, HH & Simpson, G 1997, 'Capabilities and Limitations of Ion Mobility Spectrometry for Field Screening Applications', *Annual Review of Physical Chemistry*, vol. 49, pp. 203-232.

Hinze, J 1975, *Turbulence (2nd edn)*, McGraw-Hill: New York.

Hopkins, AR & Lewis, NS 2001, 'Detection and classification characteristics of arrays of carbon black/organic polymer composite chemiresistive vapor detectors for the nerve agent

simulants dimethylmethylphosphonate and diisopropylmethylphosponate', *Analytical Chemistry*, vol. 73, no. 5, pp. 884-892.

Hu, C-Y & Raymond, DJ 2004, 'Lessons learned from hazardous chemical incidentsâ€"Louisiana Hazardous Substances Emergency Events Surveillance (HSEES) system', *Journal of Hazardous Materials*, vol. 115, no. 1â€"3, pp. 33-38.

Hyvärinen, A 1999, 'Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation', *Neural Computation*, vol. 11, no. 7, pp. 1739–1768.

Hyvärinen, A & Oja, E 2000, 'Independent Component Analysis: Algorithms and Applications', *Neural Networks*, vol. 13, no. 4-5, pp. 411-430.

Intra, P & Tippayawong, N 2008, 'An overview of differential mobility analyzers for size classification of nanometer-sized aerosol particles', *Songklanakarin Journal of Science and Technology*, vol. 30, no. 2, pp. 243-256.

IONSCAN, *http://www.smithsdetection.com/explosives_narcotics_trace.php*, Smiths Detection.

Ishida, H 2009, 'Blimp Robot for Three-Dimensional Gas Distribution Mapping in Indoor Environment', in *Olfaction and Electronic Nose, Proceedings*, vol. 1137, eds M Pardo & G Sberveglieri, Amer Inst Physics, Melville, pp. 61-64.

Ishida, H, Kagawa, Y, Nakamoto, T & Moriizumi, T 1996, 'Odor-source localization in the clean room by an autonomous mobile sensing system', *Sensors and Actuators B: Chemical*, vol. 33, no. 1-3, pp. 115-121.

Ishida, H, Kobayashi, A, Nakamoto, T & Moriizumi, T 1999, 'Three-dimensional odor compass', *Ieee Transactions on Robotics and Automation*, vol. 15, no. 2, pp. 251-257.

Ishida, H, Nakamoto, T & Moriizumi, T 1998, 'Remote sensing of gas/odor source location and concentration distribution using mobile system', *Sensors and Actuators B: Chemical*, vol. 49, no. 1-2, pp. 52-57.

Ishida, H, Nakayama, G, Nakamoto, T & Moriizumi, T 2005, 'Controlling a gas/odor plume-tracking robot based on transient responses of gas sensors', *Sensors Journal, IEEE*, vol. 5, no. 3, pp. 537-545.

Ishida, H, Suetsugu, K, Nakamoto, T & Moriizumi, T 1994, 'Study of autonomous mobile sensing system for localization of odor source using gas sensors and anemometric sensors', *Sensors and Actuators A: Physical*, vol. 45, no. 2, pp. 153-157.

Ishida, H, Tanaka, H, Taniguchi, H & Moriizumi, T 2006, 'Mobile robot navigation using vision and olfaction to search for a gas/odor source', *Autonomous Robots*, vol. 20, no. 3, pp. 231-238.

Ishida, H, Wada, Y & Matsukura, H 2012, 'Chemical Sensing in Robotic Applications: A Review', *Sensors Journal, IEEE*, vol. 12, no. 11, pp. 3163-3173.

Ishida, H, Zhu, M, Johansson, K & Moriizumi, T 2004, 'Three-Dimensional Gas/Odor Plume Tracking with Blimp', in *Proceedings ICEE*, Sapporo, Japan, pp. 117–120.

Jacobson, SH, McLay, LA, Kobza, JE & Bowman, JM 2005, 'Modeling and analyzing multiple station baggage screening security system performance', *Naval Research Logistics (NRL)*, vol. 52, no. 1, pp. 30-45.

Jakuba, MV 2007, Stochastic mapping for chemical plume source localization with application to autonomous hydrothermal vent discovery, thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution.

James, D, Scott, SM, Ali, Z & O'Hare, WT 2005, 'Chemical sensors for electronic nose systems', *Microchimica Acta*, vol. 149, no. 1-2, pp. 1-17.

Jatmiko, W, Mursanto, P, Kusumoputro, B, Sekiyama, K & Fukuda, T 2008, 'Modified PSO algorithm based on flow of wind for odor source localization problems in dynamic environments', *WSEAS Transations on systems*, vol. 7, no. 2, pp. 106-113.

Jatmiko, W, Sekiyama, K & Fukuda, T 2007, 'A PSO-based mobile robot for odor source localization in dynamic advection-diffusion with obstacles environment: theory, simulation and measurement', *IEEE Computational Intelligence Magazine*, pp. 37-51.

Jaumot, J, Gargallo, R, Juan, Ad & Tauler, R 2005, 'A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB', *Chemometrics and Intelligent Laboratory Systems*, vol. 76, pp. 101-110.

Jaumot, J & Tauler, R 2010, 'MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution', *Chemometrics and Intelligent Laboratory Systems*, vol. 103, no. 2, pp. 96-107.

Jazan, E & Tabrizchi, M 2009, 'Kinetic study of proton-bound dimer formation using ion mobility spectrometry', *Chemical Physics*, vol. 355, no. 1, pp. 37-42.

Jolliffe, IT 2002, *Principal Component Analysis*, Springer Series in Statistics, New York.

Juan, Ad, Maeder, M, Martinez, M & Tauler, R 2000, 'Combining hard and soft modelling to solve kinetic problems', *Chemometrics and Intelligent Laboratory Systems*, vol. 54, pp. 123-141.

Juan, AD, VanderHeyden, Y, Tauler, R & Massart, DL 1997, 'Assessment of new constraints applied to the alternating least squares method', *Analytica Chimica Acta*, vol. 346, pp. 307-318.

Kanu, AB, Dwivedi, P, Tam, M, Matz, L & Hill, HH 2008, 'Ion mobility–mass spectrometry', *Journal of Mass Spectrometry*, vol. 43, no. 1, pp. 1-22.

Kanu, AB, Haigh, PE & Hill, HH 2005, 'Surface detection of chemical warfare agent simulants and degradation products', *Analytica Chimica Acta*, vol. 553, pp. 148-159.

Kanu, AB & Hill Jr, HH 2008, 'Ion mobility spectrometry detection for gas chromatography', *Journal of Chromatography A*, vol. 1177, no. 1, pp. 12-27.

Kanu, AB, Wu, C & Hill, HH 2008, 'Rapid preseparation of interferences for ion mobility spectrometry', *Analytica Chimica Acta*, vol. 610, no. 1, pp. 125-134.

Keller, HR & Massart, DL 1992, 'Evolving factor analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 12, no. 3, pp. 209-224.

Keller, T, Miki, A, Regenscheit, P, Dirnhofer, R, Schneider, A & Tsuchihashi, H 1998, 'Detection of designer drugs in human hair by ion mobility spectrometry (IMS)', *Forensic Science International*, vol. 94, no. 1-2, pp. 55-63.

Kennedy, J & Eberhart, R 1995, 'Particle swarm optimization', in *Neural Networks, 1995. Proceedings., IEEE International Conference on*, IEEE, pp. 1942-1948.

Kim, YI, Yoon, DJ, Lee, SS, Lee, YH & Kim, KB 2011, 'Application of laser-induced breakdown spectroscopy for detecting leakage of boric acid', *Ndt & E International*, vol. 44, no. 3, pp. 311-314.

Knutson, EO & Whitby, KT 1975, 'Aerosol classification by electric mobility: apparatus, theory, and applications', *Journal of Aerosol Science*, vol. 6, no. 6, pp. 443-451.

Kohl, D, Kelleter, J & Petig, H 2001, 'Detection of Fires by Gas Sensors', *Sensors Update*, vol. 9, no. 1, pp. 161-223.

Komsta, L 2011, 'Comparison of Several Methods of Chromatographic Baseline Removal with a New Approach Based on Quantile Regression', *Chromatographia*, vol. 73, no. 7-8, pp. 721-731.

Kowadlo, G, Rawlinson, D, Russell, RA & Jarvis, R 2006, 'Bi-modal search using complementary sensing (olfaction/vision) for odour source localisation', in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, IEEE, pp. 2041-2046.

Kowadlo, G & Russell, RA 2006, 'Using naive physics for odor localization in a cluttered indoor environment', *Auton. Robots*, vol. 20, no. 3, pp. 215-230.

Kowadlo, G & Russell, RA 2008, 'Robot Odor Localization: A Taxonomy and Survey', *Int. J. Rob. Res.*, vol. 27, no. 8, pp. 869-894.

Krebs, MD, Kang, J-M, Cohen, SJ, Lozow, JB, Tingley, RD & Davis, CE 2006, 'Two-dimensional alignment of differential mobility spectrometer data', *Sensors and Actuators B: Chemical*, vol. 119, no. 2, pp. 475-482.

Krishnanand, K & Ghose, D 2009, 'Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions', *Swarm intelligence*, vol. 3, no. 2, pp. 87-124.

Krylov, E, Nazarov, EG, Miller, RA, Tadjikov, B & Eiceman, GA 2002, 'Field Dependence of Mobilities for Gas-Phase-Protonated Monomers and Proton-Bound Dimers of Ketones by Planar Field Asymmetric Waveform Ion Mobility Spectrometer (PFAIMS)', *The Journal of Physical Chemistry A*, vol. 106, no. 22, pp. 5437-5444.

Lai, H, Corbin, I & Almirall, JR 2008, 'Headspace sampling and detection of cocaine, MDMA, and marijuana via volatile markers in the presence of potential interferences by solid phase microextraction-ion mobility spectrometry (SPME-IMS)', *Analytical and Bioanalytical Chemistry*, vol. 392, no. 1-2, pp. 105-113.

Lai, H, Guerra, P, Joshi, M & Almirall, JR 2008, 'Analysis of volatile components of drugs and explosives by solid phase microextraction-ion mobility spectrometry', *Journal of Separation Science*, vol. 31, no. 2, pp. 402-412.

Larionova, S, Almeida, N, Marques, L & Almeida, AT 2006, 'Olfactory coordinated area coverage', *Auton. Robots*, vol. 20, no. 3, pp. 251-260.

Lawrence, AH 1987, 'Detection of drugs residues on the hands of subjects by surface sampling and Ion Mobility Spectrometry', *Forensic Science International*, vol. 34, no. 1-2, pp. 73-83.

References

Lawrence, AH, Barbour, RJ & Sutcliffe, R 1991, 'Identification of wood species by ion mobility spectrometry', *Analytical Chemistry*, vol. 63, no. 13, pp. 1217-1221.

Leahy-Hoppa, M, Fitch, M & Osiander, R 2009, 'Terahertz spectroscopy techniques for explosives detection', *Analytical and Bioanalytical Chemistry*, vol. 395, no. 2, pp. 247-257.

Li, F, Xie, Z, Schmidt, H, Sielemann, S & Baumbach, JI 2002, 'Ion mobility spectrometer for online monitoring of trace compounds', *Spectrochimica Acta Part B*, vol. 57, pp. 1563–1574.

Li, J-G, Meng, Q-H, Wang, Y & Zeng, M 2011, 'Odor source localization using a mobile robot in outdoor airflow environments with a particle filter algorithm', *Autonomous Robots*, vol. 30, no. 3, pp. 281-292.

Lilienthal, A & Duckett, T 2004, 'Building gas concentration gridmaps with a mobile robot', *Robotics and Autonomous Systems*, vol. 48, pp. 3-16.

Lilienthal, A, Loutfi, A & Duckett, T 2006, 'Airborne Chemical Sensing with Mobile Robots', *Sensors*, vol. 6, no. 11, pp. 1616-1678.

Lilienthal, A, Reimann, D & Zell, A 2003, 'Gas Source Tracing with a Mobile Robot Using an Adapted Moth Strategy', in *Proceedings of Autonomous Mobile Systems*.

Lilienthal, A, Streichert, F & Zell, A 2005, 'Model-based shape analysis of gas concentration gridmaps for improved gas source localisation', in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, IEEE, pp. 3564-3569.

Lilienthal, A, Ulmer, H, Frohlich, H, Werner, F & Zell, A 2004, 'Learning to detect proximity to a gas source with a mobile robot', in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, pp. 1444-1449 vol.2.

Lilienthal, A, Zell, A, Wandel, M & Weimar, U 2001, 'Sensing odour sources in indoor environments without a constant airflow by a mobile robot', in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, pp. 4005-4010 vol.4.

Lilienthal, AJ, Duckett, T, Ishida, H & Werner, F 2006, 'Indicators of Gas Source Proximity using Metal Oxide Sensors in a Turbulent Environment', in *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, pp. 733-738.

Lilienthal, AJ, Reggente, M, Trincavelli, M, Blanco, JL & Gonzalez, J 2009, 'A Statistical Approach to Gas Distribution Modelling with Mobile Robots – The Kernel DM+V Algorithm', in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 570-576.

Liu, Y & Passino, KM 2002, 'Biomimicry of Social Foraging Bacteria for Distributed Optimization: Models, Principles, and Emergent Behaviors', *Journal of Optimization Theory and Applications*, vol. 115, no. 3, pp. 603-628.

Lo Iacono, G 2010, 'A Comparison of Different Searching Strategies to Locate Sources of Odor in Turbulent Flows', *Adaptive Behavior*, vol. 18, no. 2, pp. 155-170.

Louis, M, Huber, T, Benton, R, Sakmar, TP & Vosshall, LB 2008, 'Bilateral olfactory sensory input enhances chemotaxis behavior', *Nat Neurosci*, vol. 11, no. 2, pp. 187-199.

Loutfi, A & Coradeschi, S 2006, 'Smell, think and act: A cognitive robot discriminating odours', *Autonomous Robots*, vol. 20, no. 3, pp. 239-249.

Loutfi, A, Coradeschi, S, Lilienthal, AJ & Gonzalez, J 2009, 'Gas distribution mapping of multiple odour sources using a mobile robot', *Robotica*, vol. 27, no. 2, pp. 311-319.

Lu, Y, O'Donnell, RM & Harrington, PB 2009, 'Detection of cocaine and its metabolites in urine using solid phase extraction-ion mobility spectrometry with alternating least squares', *Forensic Science International*, vol. 189, no. 1-3, pp. 54-59.

Lytridis, C, Virk, GS & Kadar, EE 2005, 'Co-Operative Smell-Based Navigation for Mobile Robots', in *Climbing and Walking Robots*, Springer Berlin Heidelberg, pp. 1107-1117.

Marco, S & Gutierrez-Galvez, A 2012, 'Signal and Data Processing for Machine Olfaction and Chemical Sensing: A Review', *Sensors Journal, IEEE*, vol. 12, no. 11, pp. 3189-3214.

Marjovi, A & Marques, L 2011, 'Multi-robot olfactory search in structured environments', *Robotics and Autonomous Systems*, vol. 59, no. 11, pp. 867-881.

Marques, L, Nunes, U & Almeida, ATd 2002, 'Olfaction-based mobile robot navigation', *Thin Solid Films*, vol. 418, pp. 51-58.

Marques, L, Nunes, U & Almeida, ATd 2006, 'Particle swarm-based olfactory guided search', *Auton. Robots*, vol. 20, pp. 277-287.

Márquez-Sillero, I, Aguilera-Herrador, E, Cárdenas, S & Valcárcel, M 2011, 'Ion-mobility spectrometry for environmental analysis', *TrAC Trends in Analytical Chemistry*, vol. 30, no. 5, pp. 677-690.

Martinez, D & Perrinet, L 2002, 'Cooperation between vision and olfaction in a koala robot', *Report on the 2002 Workshop on Neuromorphic Engineering*, pp. 51-53.

Matz, LM, Tornatore, PS & Hill, HH 2001, 'Evaluation of suspected interferents for TNT detection by ion mobility spectrometry', *Talanta*, vol. 54, no. 1, pp. 171-179.

Mcgill, K & Taylor, S 2011, 'Robot algorithms for localization of multiple emission sources', *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1-25.

Meixner, H & Lampe, U 1996, 'Metal oxide sensors', *Sensors and Actuators B: Chemical*, vol. 33, no. 1-3, pp. 198-202.

MiniRAE3000, *http://www.raesystems.com/products/minirae-3000*, RAE systems.

Mitzner, KD, Sternhagen, J & Galipeau, DW 2003, 'Development of a micromachined hazardous gas sensor array', *Sensors and Actuators B-Chemical*, vol. 93, no. 1-3, pp. 92-99.

Moore, D 2007, 'Recent Advances in Trace Explosives Detection Instrumentation', *Sensing and Imaging: An International Journal*, vol. 8, no. 1, pp. 9-38.

MultiRAEplus, *http://www.raesystems.com/products/multirae-plus*, RAE systems.

Murlis, J, Elkinton, JS & Carde, RT 1992, 'Odor plumes and how insects use them', *Annual Review of Entomology*, no. 37, pp. 505-532.

Mylne, KR & Mason, PJ 1991, 'Concentration fluctuation measurements in a dispersing plume at a range of up to 1000m', *Q. J. R. Meteorol. Soc.*, vol. 117, pp. 177-206.

References

Na, N, Zhang, C, Zhao, MX, Zhang, SC, Yang, CD, Fang, X & Zhang, XR 2007, 'Direct detection of explosives on solid surfaces by mass spectrometry with an ambient ion source based on dielectric barrier discharge', *Journal of Mass Spectrometry*, vol. 42, no. 8, pp. 1079-1085.

Nagle, HT, Gutierrez-Osuna, R & Schiffman, SS 1998, 'The how and why of electronic noses', *Ieee Spectrum*, vol. 35, no. 9, pp. 22-34.

Nakamoto, T, Ishida, H & Moriizumi, T 1996, 'An odor compass for localizing an odor source', *Sensors and Actuators B: Chemical*, vol. 35, no. 1-3, pp. 32-36.

O' Donnell, RM, Sun, X & Harrington, PdB 2008, 'Pharmaceutical applications of ion mobility spectrometry', *TrAC Trends in Analytical Chemistry*, vol. 27, no. 1, pp. 44-53.

O'Hara, MP 2008, Detection of IED emplacement in urban environments, thesis, Monterey, California. Naval Postgraduate School.

Ochoa, ML & Harrington, PB 2004, 'Detection of Methamphetamine in the Presence of Nicotine Using In Situ Chemical Derivatization and Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 76, no. 4, pp. 985-991.

Ochoa, ML & Harrington, PB 2005, 'Chemometric studies for the characterization and differentiation of microorganisms using in situ derivatization and thermal desorption ion mobility spectrometry', *Analytical Chemistry*, vol. 77, no. 3, pp. 854-863.

Ogden, ID & Strachan, NJC 1993, 'Enumeration of escherichia-coli in cooked and raw meats by ion mobility spectrometry', *Journal of Applied Bacteriology*, vol. 74, no. 4, pp. 402-405.

Olfati-Saber, R 2007, 'Distributed tracking for mobile sensor networks with information-driven mobility', in *American Control Conference, 2007. ACC'07*, IEEE, pp. 4606-4612.

Pang, S & Farrell, JA 2006, 'Chemical Plume Source Localization', *IEEE Transactions on systems, man, and cybernetics*, vol. 36, no. 5, pp. 1068-1080.

PEN3, *http://www.airsense.com/en/products/pen3/*, Airsense Analytics.

Peng, J, Peng, S, Jiang, A, Wei, J, Li, C & Tan, J 2010, 'Asymmetric least squares for multiple spectra baseline correction', *Analytica Chimica Acta*, vol. 683, no. 1, pp. 63-68.

Perera, A, Papamichail, N, Barsan, N, Weimar, U & Marco, S 2006, 'On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions', *Sensors Journal, IEEE*, vol. 6, no. 3, pp. 770-783.

Persaud, K & Dodd, G 1982, 'Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose', *Nature*, vol. 299, no. 5881, pp. 352-355.

Purnamadjaja, AH & Russell, RA 2005, 'Pheromone communication in a robot swarm: necrophoric bee behaviour and its replication', *Robotica*, vol. 23, no. 6, pp. 731-742.

Puton, J, Holopainen, SI, Makinen, MA & Sillanpaa, MET 2012, 'Quantitative Response of IMS Detector for Mixtures Containing Two Active Components', *Analytical Chemistry*, vol. 84, no. 21, pp. 9131-9138.

Pyk, P, Bermúdez i Badia, S, Bernardet, U, Knüsel, P, Carlsson, M, Gu, J, Chanie, E, Hansson, B, Pearce, T & J. Verschure, P 2006, 'An artificial moth: Chemical source localization using a robot based neuronal model of moth optomotor anemotactic search', *Autonomous Robots*, vol. 20, no. 3, pp. 197-213.

Ramaker, HJ, van Sprang, ENM, Westerhuis, JA & Smilde, AK 2003, 'Dynamic time warping of spectroscopic BATCH data', *Analytica Chimica Acta*, vol. 498, no. 1-2, pp. 133-153.

RAMEM, *http://www.ioner.eu/*, RAMEM S.A.

Raub, JA, Mathieu-Nolf, M, Hampson, NB & Thom, SR 2000, 'Carbon monoxide poisoning - A public health perspective', *Toxicology*, vol. 145, no. 1, pp. 1-14.

Rearden, P & Harrington, PB 2005, 'Rapid screening of precursor and degradation products of chemical warfare agents in soil by solid-phase microextraction ion mobility spectrometry (SPME–IMS)', *Analytica Chimica Acta*, vol. 545, pp. 13-20.

Reggente, M & Lilienthal, AJ 2010, 'The 3D-Kernel DM+V/W Algorithm: Using Wind Information in Three Dimensional Gas Distribution Modelling with a Mobile Robot', in *Proceedings of IEEE Sensors*, pp. 999-1004.

Reggente, M, Mondini, A, Ferri, G, Mazzolai, B, Gabelletti, M, Dario, P & Lilienthal, AJ 2010, 'The DustBot System: Using Mobile Robots to Monitor Pollution in Pedestrian Area', *Chemical Engineering Transactions*, vol. 23, pp. 273-278.

Rock, F, Barsan, N & Weimar, U 2008, 'Electronic nose: Current status and future trends', *Chemical Reviews*, vol. 108, no. 2, pp. 705-725.

Rosipal, R 2008, 'Nonlinear Partial Least Squares An Overview', in *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pp. 169-189.

Rozas, R, Morales, J & Vega, D 1991, 'Artificial smell detection for robotic navigation', in *Advanced Robotics, 1991. 'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*, pp. 1730-1733 vol.2.

RQbox, *http://env.alpha-mos.com/en/produits/rq_box.php*, Alpha MOS.

Russell, RA 1999, 'Ant trails - an example for robots to follow?', in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, pp. 2698-2703 vol.4.

Russell, RA 2001, 'Tracking chemical plumes in constrained environments', *Robotica*, vol. 19, no. 4, pp. 451-458.

Russell, RA 2004a, 'Locating underground chemical sources by tracking chemical gradients in 3 dimensions', in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, pp. 325-330 vol.1.

Russell, RA 2004b, 'Robotic location of underground chemical sources', *Robotica*, vol. 22, no. 1, pp. 109-115.

Russell, RA 2011, 'CRABOT: A Biomimetic Burrowing Robot Designed for Underground Chemical Source Location', *Advanced Robotics*, vol. 25, no. 1-2, pp. 119-134.

Russell, RA, Bab-Hadiashar, A, Shepherd, RL & Wallace, GG 2003, 'A comparison of reactive robot chemotaxis algorithms', *Robotics and Autonomous Systems*, vol. 45, no. 2, pp. 83-97.

Russell, RA, Thiel, D, Deveza, R & Mackay-Sim, A 1995, 'A robotic system to locate hazardous chemical leaks', in *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, pp. 556-561 vol.1.

Ruzsanyi, V & Baumbach, J 2005, 'Analysis of human breath using IMS', *International Journal for Ion Mobility Spectrometry*, vol. 8, pp. 5-7.

Ruzsanyi, V, Baumbach, JrI, Sielemann, S, Litterst, P, Westhoff, M & Freitag, L 2005, 'Detection of human metabolites using multi-capillary columns coupled to ion mobility spectrometers', *Journal of Chromatography A*, vol. 1084, no. 1-2, pp. 145-151.

Sabo, M & Matejčík, Š 2012, 'Corona Discharge Ion Mobility Spectrometry with Orthogonal Acceleration Time of Flight Mass Spectrometry for Monitoring of Volatile Organic Compounds', *Analytical Chemistry*, vol. 84, no. 12, pp. 5327-5334.

Sabo, M, Matúška, J & Matejčík, Š 2011, 'Specific O2 generation in corona discharge for ion mobility spectrometry', *Talanta*, vol. 85, no. 1, pp. 400-405.

Sandini, G, Lucarini, G & Varoli, M 1993, 'Gradient driven self-organizing systems', in *Intelligent Robots and Systems '93, IROS '93. Proceedings of the 1993 IEEE/RSJ International Conference on*, pp. 429-432 vol.1.

Santos, JP, Hontanon, E, Ramiro, E & Alonso, M 2009, 'Performance evaluation of a high-resolution parallel-plate differential mobility analyzer', *Atmospheric Chemistry and Physics*, vol. 9, no. 7, pp. 2419-2429.

Savitzky, A & Golay, MJE 1964, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', *Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639.

Savorani, F, Tomasi, G & Engelsen, SB 2010, 'icoshift: A versatile tool for the rapid alignment of 1D NMR spectra', *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190-202.

Scott, SM, James, D & Ali, Z 2006, 'Data analysis for electronic nose systems', *Microchimica Acta*, vol. 156, no. 3-4, pp. 183-207.

Schweizerberberich, PM, Vaihinger, S & Gopel, W 1994, 'Characterization of Food Freshness with sensor arrays', *Sensors and Actuators B-Chemical*, vol. 18, no. 1-3, pp. 282-290.

SentinelII, *http://www.smithsdetection.com/esp/1522.php*, Smiths Detection.

Sferopoulos, R 2009, 'A Review of Chemical Warfare Agent (CWA) Detector Technologies and Commercial-Off-The-Shelf Items', *Defence Science and Technology Organisation*, vol. Department of Defence, Australian Government.

Shamlouei, HR & Tabrizchi, M 2008, 'Transmission of different ions through a drift tube', *International Journal of Mass Spectrometry*, vol. 273, no. 1-2, pp. 78-83.

Shao, XG, Leung, AKM & Chau, FT 2003, 'Wavelet: A new trend in chemistry', *Accounts of Chemical Research*, vol. 36, no. 4, pp. 276-283.

Shea, P, Gozani, T & Bozorgmanesh, H 1990, 'A TNA explosives-detection system in airline baggage', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 299, no. 1-3, pp. 444-448.

Shvartsburg, AA & Smith, RD 2008, 'Fundamentals of Traveling Wave Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 80, no. 24, pp. 9689-9699.

Simpson, G, Klasmeier, M, Hill, H, Atkinson, D, Radolovich, G, Lopez-Avila, V & Jones, TL 1996, 'Evaluation of gas chromatography coupled with ion mobility spectrometry for

monitoring vinyl chloride and other chlorinated and aromatic compounds in air samples', *Journal of High Resolution Chromatography*, vol. 19, no. 6, pp. 301-312.

Singh, S & Singh, M 2003, 'Explosives detection systems (EDS) for aviation security', *Signal Processing*, vol. 83, pp. 31-55.

Skov, T, van den Berg, F, Tomasi, G & Bro, R 2006, 'Automated alignment of chromatographic data', *Journal of Chemometrics*, vol. 20, no. 11-12, pp. 484-497.

Snyder, AP, Blyth, DA & Parsons, JA 1996, 'Ion mobility spectrometry as an immunoassay detection technique', *Journal of Microbiological Methods*, vol. 27, no. 1, pp. 81-88.

Spangler, GE 2002, 'Expanded theory for the resolving power of a linear ion mobility spectrometer', *International Journal of Mass Spectrometry*, vol. 220, pp. 399-418.

Stachniss, C, Plagemann, C & Lilienthal, AJ 2009, 'Learning gas distribution models using sparse Gaussian process mixtures', *Auton. Robots*, vol. 26, no. 2-3, pp. 187-202.

Stone, M & Brooks, RJ 1990, 'Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 52, no. 2, pp. 237-269.

Strachan, NJC, Nicholson, FJ & Ogden, ID 1995, 'An automated sampling system using ion mobility spectrometry for the rapid detection of bacteria', *Analytica Chimica Acta*, vol. 313, no. 1-2, pp. 63-67.

Sutton, OG 1947, 'The problem of diffusion in the lower atmosphere', *Q. J. R. Meteorol. Soc.*, vol. 73, pp. 257-281.

Tauler, R 1995, 'Multivariate curve resolution applied to second order data', *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 133-146.

Tauler, R, Kowalski, B & Fleming, S 1993, 'Multivariate curve resolution applied to spectral data from multiple runs of an industrial process', *Analytical Chemistry*, vol. 65, no. 15, pp. 2040-2047.

Tauler, R, Smilde, A & Kowalski, B 1995, 'Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution', *Journal of Chemometrics*, vol. 9, no. 1, pp. 31-58.

Tomasi, G, Savorani, F & Engelsen, SB 2011, 'icoshift: An effective tool for the alignment of chromatographic data', *Journal of Chromatography A*, vol. 1218, no. 43, pp. 7832-7840.

Tomasi, G, van den Berg, F & Andersson, C 2004, 'Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data', *Journal of Chemometrics*, vol. 18, no. 5, pp. 231-241.

Trincavelli, M 2010, Gas discrimination for mobile robots, thesis, Örebro universitet.

Trincavelli, M, Coradeschi, S & Loutfi, A 2009, 'Odour classification system for continuous monitoring applications', *Sensors and Actuators B-Chemical*, vol. 139, no. 2, pp. 265-273.

Trincavelli, M, Reggente, M, Coradeschi, S, Loutfi, A, Ishida, H & Lilienthal, AJ 2008, *Towards Environmental Monitoring with Mobile Robots*, Ieee, New York.

References

Tuovinen, K, Paakkanen, H & Hänninen, O 2000, 'Detection of pesticides from liquid matrices by ion mobility spectrometry', *Analytica Chimica Acta*, vol. 404, no. 1, pp. 7-17.

Utriainen, M, Kärpänoja, E & Paakkanen, H 2003, 'Combining miniaturized ion mobility spectrometer and metal oxide gas sensor for the fast detection of toxic chemical vapors', *Sensors and Actuators B*, vol. 93, pp. 17-24.

Vanneste, E & Geise, HJ 2004, 'Commercial Electronic Nose Instruments', in *Handbook of Machine Olfaction*, Wiley-VCH Verlag GmbH & Co. KGaA, pp. 161-179.

Vautz, W, Sielemann, S & Baumbach, JI 2004, 'Determination of terpenes in humid ambient air using ultraviolet ion mobility spectrometry', *Analytica Chimica Acta*, vol. 513, no. 2, pp. 393-399.

Vautz, W, Zimmermann, D, Hartmann, M, Baumbach, JI, Nolte, J & Jung, J 2006, 'Ion mobility spectrometry for food quality and safety', *Food Additives and Contaminants*, vol. 23, no. 11, pp. 1064-1073.

Vergassola, M, Villermaux, E & Shraiman, BI 2007, 'Infotaxis as a strategy for searching without gradients', *Nature*, vol. 445, no. 7126, pp. 406-409.

Vidal-de-Miguel, G, Macía, M & Cuevas, J 2012, 'Transversal Modulation Ion Mobility Spectrometry (TM-IMS), A New Mobility Filter Overcoming Turbulence Related Limitations', *Analytical Chemistry*, vol. 84, no. 18, pp. 7831-7837.

Vigário, R 1997, 'Extraction of ocular artifacts from EEG using independent component analysis', *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 3, pp. 395-404.

Vigário, R, Jousmäki, V, Hämäläinen, M, Hari, R & Oja, E 1998, 'Independent component analysis for identification of artifacts in magnetoencephalographic recordings', *Advances in Neural Information Processing Systems*, vol. 10, pp. 229–235.

Vigneau, E, Devaux, MF, Qannari, EM & Robert, P 1997, 'Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration', *Journal of Chemometrics*, vol. 11, no. 3, pp. 239-249.

Wallin, S, Pettersson, A, Östmark, H & Hobro, A 2009, 'Laser-based standoff detection of explosives: a critical review', *Analytical and Bioanalytical Chemistry*, vol. 395, no. 2, pp. 259-274.

Webster, DR, Volyanskyy, KY & Weissburg, MJ 2012, 'Bioinspired algorithm for autonomous sensor-driven guidance in turbulent chemical plumes', *Bioinspiration & Biomimetics*, vol. 7, no. 3, p. 11.

Webster, DR & Weissburg, MJ 2001, 'Chemosensory guidance cues in a turbulent chemical odor plume', *Limnol. Oceanogr.*, vol. 46, no. 5, pp. 1034-1047.

Westhoff, M, Freitag, PLL, Ruzsanyi, V, Bader, S, Urfer, W & Baumbach, JI 2005, 'Ion mobility spectrometry: A new method for the detection of lung cancer and airway infection in exhaled air? First results of a pilot study', *Chest*, vol. 128, no. 4, pp. 155S-155S.

Westhoff, M, Litterst, P, Freitag, L & Baumbach, J 2007, 'Ion mobility spectrometry in the diagnosis of sarcoidosis: results of a feasibility study', *J Physiol Pharmacol.*, vol. 58, no. Suppl 5 (Pt 2), pp. 739-51.

Windig, W, Gallagher, NB, Shaver, JM & Wise, BM 2005, 'A new approach for interactive self-modeling mixture analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, pp. 85-96.

Windig, W & Guilment, J 1991, 'Interactive Self-Modeling Mixture Analysis', *Analytical Chemistry*, vol. 63, no. 14, pp. 1425-1432.

Wise, BM, Gallagher, NB, Bro, R, Shaver, JM, Windig, W & Koch, RS 2006, 'Multivariate Regression', in *Chemometrics tutorial*, Eigenvector Research, Inc.

Wittmer, D, Luckenbill, BK, Hill, HH & Chen, YH 1994, 'Electrospray-Ionization Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 66, no. 14, pp. 2348-2355.

Wold, S 1992, 'Nonlinear partial least squares modelling II. Spline inner relation', *Chemometrics and Intelligent Laboratory Systems*, vol. 14, no. 1-3, pp. 71-84.

Wold, S, Esbensen, K & Geladi, P 1987, 'Principal Component Analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52.

Wold, S, Kettaneh-Wold, N & Skagerberg, B 1989, 'Nonlinear PLS modeling', *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 53-65.

Wold, S, Sjöström, M & Eriksson, L 2001, 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 109-130.

Workman, JJ, Mobley, PR, Kowalski, BR & Bro, R 1996, 'Review of Chemometrics Applied to Spectroscopy: 1985-95, Part I', *Applied Spectroscopy reviews*, vol. 31, no. 1-2, pp. 73-124.

Yang, H, Griffiths, PR & Tate, JD 2003, 'Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra', *Analytica Chimica Acta*, vol. 489, no. 2, pp. 125-136.

Yee, E 2008, 'Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference', *Boundary-Layer Meteorology*, vol. 127, pp. 359-394.

Yee, E 2009, 'Probability law of concentration in plumes dispersing in an urban area', *Environmental Fluid Mechanics*, vol. 9, pp. 389-407.

Yee, E & Biltoft, CA 2004, 'Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles', *Boundary-Layer Meteorology*, vol. 111, pp. 363-415.

Yee, E & Chan, R 1997, 'A simple model for the probability density function of concentration fluctuations in atmospheric plumes', *Atmospheric Environment*, vol. 31, pp. 991-1002.

Yee, E, Wang, B-C & Lien, F-S 2009, 'Probabilistic model for concentration fluctuations in compact-source plumes in an urbant environment', *Boundary-Layer Meteorology*, vol. 130, pp. 169-208.

Yinon, J 2003, 'Peer Reviewed: Detection of Explosives by Electronic Noses', *Analytical Chemistry*, vol. 75, no. 5, pp. 98 A-105 A.

Young, D, Thomas, CLP, Breach, J, Brittain, AH & Eiceman, GA 1999, 'Extending the concentration and linear dynamic range of ion mobility spectrometry with a sheath flow inlet', *Analytica Chimica Acta*, vol. 381, no. 1, pp. 69-83.

References

Young, GO 2005, *Independent Component Analysis* in Encyclopedia of Statistics in Behavioral Science, vol. 2, John Wiley & Sons, Chinchester, pp. 907-912.

Zamora, D & Blanco, M 2012, 'Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration', *Analytica Chimica Acta*, vol. 726, no. 0, pp. 50-56.

Zarzhitsky, D, Spears, DF & Spears, WM 2005, 'Swarms for chemical plume tracing', in *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, pp. 249-256.

Zhang, Q, Sobelman, GE & He, T 2009, 'Gradient-based target localization in robotic sensor networks', *Pervasive and Mobile Computing*, vol. 5, no. 1, pp. 37-48.

Zhang, ZM, Chen, S & Liang, YZ 2010, 'Baseline correction using adaptive iteratively reweighted penalized least squares', *Analyst*, vol. 135, no. 5, pp. 1138-1146.

Zheng, P, Harrington, PdB & Davis, DM 1996, 'Quantitative analysis of volatile organic compounds using ion mobility spectrometry and cascade correlation neural networks', *Chemometrics and Intelligent Laboratory Systems*, vol. 33, pp. 121-132.

Zohora, SE, Khan, AM & Hundewale, N 2013, 'Chemical Sensors Employed in Electronic Noses: A Review', in *Advances in Computing and Information Technology, Vol 3*, vol. 178, eds N Meghanathan, D Nagamalai & N Chaki, Springer-Verlag Berlin, Berlin, pp. 177-184.

Zou, Y & Chakrabarty, K 2007, 'Distributed mobility management for target tracking in mobile sensor networks', *Mobile Computing, IEEE Transactions on*, vol. 6, no. 8, pp. 872-887.

Chapter 2

Objectives

The first objective of the present thesis is providing signal processing approaches to ameliorate the detection, identification and quantitation of analytes in the presence of interfering substances when using mobility based chemical analyzers, such as Ion Mobility spectrometers (IMS) and Differential Mobility Analyzers (DMA).

In this thesis we are going to explore the capabilities of multivariate signal and data processing techniques for mobility spectra through particular applications.

Among the myriad of techniques that can be applied to IMS spectra we are especially interested in Multivariate Curve Resolution (MCR) techniques since they are able to provide dimensionality reduction with interpretable bilinear decompositions.

Since IMS spectra are non-linear as substance concentration increases, this thesis explores how to circumvent this potential problem and combining with Partial Least Squares (PLS) the chemical concentrations of the target substances are obtained.

DMA is a recent configuration for the analysis of mobility spectra. The analysis of DMA spectra by multivariate techniques has not been attempted before. In this thesis, it is explored if DMA performance for qualitative and quantitative analysis can be improved by proper signal and data processing.

The second objective of the thesis is the design and implementation of a gas chemical source localization algorithm which integrates the chemical concentration measurements of the target substance from multiple sensors.

We intend that the algorithm provides a map of probability of the presence of the source in the explored arena, without explicitly performing plume tracking. Therefore, this implies that the mobile carriers (mounting gas sensors) could be intended for other tasks additionally to find the source.

In the proposed scenario, the chemical concentration is provided by IMS instruments and in consequence it could have been obtained using the proposed signal processing approaches in this thesis.

To test different algorithmic approaches we will develop a realistic chemical plume readings simulator for a set of vehicles carrying the mobility spectrometers (chemical sensors in general).

Moreover, in order to validate the results from simulations, real chemical source localization experiments were carried out in indoors and outdoors environments, with forced ventilation and turbulent flows.

The structure of the present doctoral dissertation is as follows:

- **Chapter 1** provided a general introduction. Some scenarios where chemicals have to be detected and quantified and others where the location of the gas source is unknown were described. An overview of the most relevant sensing technologies to the present thesis so as to detect and quantify chemicals was also provided. Signal processing approaches to deal with the signal of the sensors were discussed; and finally, a review of algorithms for chemical source localization was given.
- **Chapter 2** summarizes the objectives of the thesis.
- **Chapter 3** presents the results of applying multivariate data techniques for the first time to DMA. The performance for explosive detection, and classification and quantitation of volatile organic compounds (VOCs) is assessed.
- **Chapter 4** presents the qualitative results obtained from a new technique based on MCR. The data were obtained using IMS in a baggage security checkpoint where interfering chemicals appear.
- **Chapter 5** provides a new methodology to deal with non-linearities in IMS spectra. Qualitative and quantitative results are analyzed.
- **Chapter 6** presents a gas chemical source localization algorithm which integrates concentration information. Simulated and real-world data are used in order to test and compare this new approach with an existing algorithm based on binary detections.
- **Chapter 7** summarizes the more relevant conclusions of this doctoral dissertation.
- **Chapter 8** summarizes the thesis in Spanish.
- **Chapter 9** lists the publications and participations in conferences.

Chapter 3

Multivariate techniques applied to a novel Differential Mobility Analyzer (DMA) for explosives detection and VOCs identification and quantitation

3.1 Introduction

In most of the applications, IMS or DMA instruments are used only as qualitative detectors and only the absence or presence of the target substance is of interest. Quantitative determination of analytes with IMS technologies is typically univariate, that is, a resolved peak area or height is related to a specific concentration. However, this approach fails catastrophically when interfering chemicals produce overlapped peaks due to the limited selectivity of mobility-based technologies.

Although several papers have reported the use of multivariate signal processing to analyze either qualitative (Rauch, Harrington & Davis 1998; Reese & Harrington 1999; Ochoa & Harrington 2005; Bota & Harrington 2006; Prasad et al. 2008; Pomareda et al. 2010; Karpas et al. 2012) or quantitative information (Boger & Karpas 1994; Zheng, Harrington & Davis 1996; Fraga, Kerr & Atkinson 2009; Zamora & Blanco 2012; Pomareda et al. 2012) from IMS data, to the best of our knowledge there is a complete lack of contributions in multivariate data analysis using DMA instruments.

This chapter aims to demonstrate that explosive detection (section 3.5) and identification and quantitation of VOCs (section 3.6) with a relatively similar chemical structure are feasible with a novel Differential Mobility Analyzer. For the first time, multivariate data analysis tools are applied to the spectra obtained from a novel Differential Mobility Analyzer (DMA).

3.2 Instrumentation: DMA specifications

A fully integrated DMA prototype developed by RAMEM S.A. (Madrid, Spain) (RAMEM) has been used for all the measurements. The configuration corresponds to a planar parallel plate. The working principle of this kind of DMA is described in chapter 1 (section 1.3.2) and a scheme is shown in Figure 1.10.

The classification region has been manufactured following the most advanced machining techniques of 3D Laser sintering. The sheath gas flow is propelled by a homemade compressor and the flow rate is controlled by a turbine flow meter, model FT-16, from FTI Flow Technology (http://www.ftimeters.com/). The flow rates can range from 100L·min^{-1} until 800L·min^{-1}. The dimensions of the classification region are 5mm length and 5mm wide and the theoretical contribution to the Resolving Power (RP) achievable by the ratio between flows (sample flow and sheath flow) is as high as 250 for reduced mobilities of 0.90cm^{2}V^{-1}s^{-1}. Temperature and pressure are measured using thermocouples (type T) and pressure sensors (TCDirect model 716-090), and when necessary, the sheath flow is heated with an electric resistance of nicron to reach the desired temperature. While the instrument allows heating the sheath air up to 90ºC, data shown in this work have been recorded with the sheath air at room temperature. The ion current is determined by an electrometer developed by RAMEM to measure ultra-low currents down to 1fA (EL-5020). Sample introduction and exhaust flow rates through the inlet and outlet slits are controlled by PID loops in order to control quantitative response. Ionization of the analytes has been done using a Heraeus UV lamp with energy of 10.6eV, located as close as possible to the inlet of the classification region. All the components are controlled by an OPLC (Open Programmable Logic Computer) from Unitronics V1040 and with specific software developed using LabView 2009 (National Instruments, Austin, Texas, USA) that also includes the data processing capabilities.
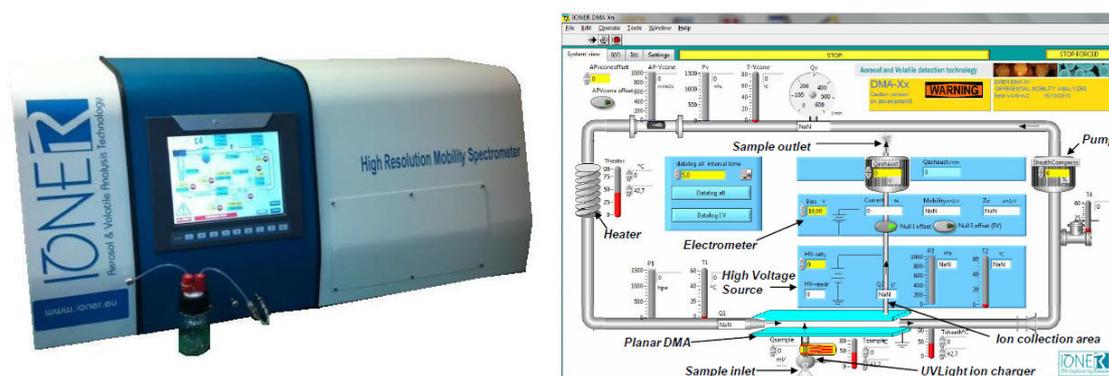
Figure 3.1. Picture of the High Resolution Mobility Spectrometer (HRIMS) based on DMA manufactured by RAMEM and its graphical interface for instrument control, data acquisition, data processing and data visualization.

The time needed to complete one spectrum is given by the time it takes to the High Voltage source to change from the starting point to the end point of the voltage; this is normally less than 1 minute. The voltage, where certain peak appears, increases with

the sheath flow rate. A maximum sheath flow rate of 600L·min^{-1} is achieved in the current model of the DMA and maximum voltages of 6KV are allowed by the electrode configuration.

The measurements were done using low sheath flow rates to keep the sensitivity, as low as 160L·min^{-1}. The resulting voltages ranged from 1.0kV to 4.0kV. Theses values allow classification of molecules with mobilities ranging from 0.2 to 3cm^{2}V^{-1}s^{-1}. The formula for the calculation of the mobility can be seen in (Eq. 3.1):

(Eq. 3.1)
$$K = \frac{Q}{V} \frac{\Delta y}{\Delta x \Delta z}$$

Where *Q* is the sheath flow-rate, *V* is the applied voltage and *Δy*, *Δx* and *Δz* are the three dimensions of the classification region. The DMA has been tested in RAMEM (Madrid, Spain) with a mobility standard like TetraHeptyl tetraAmonium Bromide (THABr), which was ionized using ElectroSpray Ionization (ESI).

The maximum value of the RP reached for the presented DMA at low sheath flow rate (300L·min^{-1}) is 45. Further experiments are being done with higher sheath flow rates for a new prototype taking special care of the mentioned problems that can be found in DMAs (Vidal-de-Miguel, Macía & Cuevas 2012). However, the obtained RP seemed good enough to differentiate among the substances presented in this chapter.

3.3 DMA software

The DMA operated by means of software implemented in LabView (National Instruments) which provided instrument configuration and control, data acquisition and data visualization. Later on, a library of functions to process DMA spectra was developed (by the author of this thesis) and integrated within the software. This library provides new functionalities for DMA spectra analysis. Some pre-processing, multivariate signal processing approaches and diagnostic tools have been implemented.

Within pre-processing methods, several functions have been included: interpolation, baseline removal, filtering, and spectra normalization (see section 3.4.1).

Within multivariate data processing approaches: PCA, LDA and PLS are included. Moreover, the KNN classifier has also been implemented.

The library also incorporates different diagnostic tools:
- Suggestions about the optimum number of components for building calibration models are made. While in the case of PCA the suggestion is based on the explained variance by each principal component, in the case of PLS this suggestion is based on different cross-validation procedures (random selection, k-fold and

leave-one-out). In the case of LDA, the number of components is decided to be the number of classes minus one.

- Based on a confidence level defined by the user (95% by default), Q (residuals) and Hotelling's T^2 statistics, abnormal samples are labelled as outliers and they should be removed before building the final calibration model.

- After substance classification using a KNN classifier, a confidence value based on the Mahalanobis' distances to the centroids of each class (substance) is assigned to each predicted/unknown sample. The distance of Mahalanobis (d_{ij}) is defined as:

(Eq. 3.2)
$$d_{ij} = \sqrt{(x_i - x_j)C_n \cdot (x_i - x_j)^T}$$

where x is a sample in a n-dimensional space, C_n is the n-dimensional covariance matrix.

The confidence is defined as:

(Eq. 3.3)
$$conf_k = \frac{1/d_k}{\sum_{i=1}^{N} 1/d_i}$$

where k is the sub-index related to the assigned class to the predicted/unknown sample and N is the total number of classes. From the last expression, it is seen that when the Mahalanobis distance d_k tends to 0, the confidence in the class k tends to 1. When d_k is exactly 0, in order to avoid a numerical problem, the confidence is forced to be 1, which is what one would expect.

- Additionally, the user can define a threshold on the confidence of classification below which predicted/unknown samples are classified as "none".

3.4 DMA spectra analysis

3.4.1 Spectra pre-processing

The first step in the pre-processing of DMA datasets consists in the homogenization of the spectra in terms of reduced mobilities (K_0) (Eq. 1.2), which are defined as the mobilities in standard conditions (20ºC and 1 bar) according to the Environmental Protection Agency (EPA) and the National Institute of Standards and Technology (NIST). In order to have homogeneous, uniform sampling in the reduced mobility axis,

a linear interpolation of the spectra is performed. In order to remove mathematical artefacts due to signal acquisition problems and reduce noise, a median filter (narrow enough to ensure no deformations in the signal) and a Savizky-Golay filter (Savitzky & Golay 1964) with a polynomial order of 3 and a window of 11 are applied. Afterwards, the baseline of each spectrum is removed independently by applying the automatic thresholding method described in (Gan, Ruan & Mo 2006). In qualitative analysis, after interpolation, filtering and baseline removal, each spectrum was normalized to unit area. Finally, only the richest part of the spectra was used, which corresponds to values of K_0 between 0.84cm^2/(V·s) and 2.14 cm^2/(V·s).

3.4.2 DMA spectra multivariate analysis

As pointed before, for the first time multivariate signal processing approaches have been applied to DMA spectra. Since the DMA is designated to operate identifying and quantifying unknown samples in real-time in field applications in real environments, calibration models must be used.

3.4.2.1 Building calibration models (off-line)

When building calibration models, training data is used. Data for training usually contains a large number of measurements (spectra) obtained under controlled conditions. These data correspond to the responses of the instrument to different substances and mixtures of substances and the known concentrations of each substance present in the sample.

Using these training data, classification models (section 1.4.2.2) based on dimensionality reduction methods such as PCA, LDA or PLS-DA can be built. On the model subspace, k-NN classifiers are used. Moreover, calibration models for substance quantitation based on PLS can also be built.

All this calibration procedure can be executed off-line. This means that all the data is previously obtained under controlled conditions, for instance in a laboratory, and the different models are built offline afterwards.

These models of reference can be used then in a real environment for prediction of unknown samples in real-time.

3.4.2.2 Prediction of unknown samples (real-time)

Data for prediction can be considered as responses obtained from the instrument under uncontrolled conditions (e.g. real environment) and it is data which needs further analysis in order to detect different substances and quantify them regarding to a library of known substances measured in a calibration step (see section 3.4.2.1).

When the instrument operates in the field in a real environment, data for prediction is produced in real-time and must be processed as it is generated (Figure 3.2).

Based on PCA, LDA or PLS models, a substance from the library of training data is assigned to unknown samples using a KNN classifier on the scores of the training model. Moreover, a confidence in substance identification is also given, and there is the possibility that a sample is classified as "none" or as "outlier" if it is very different from anything known. These unknown samples are also quantified using PLS models based on the training substances.
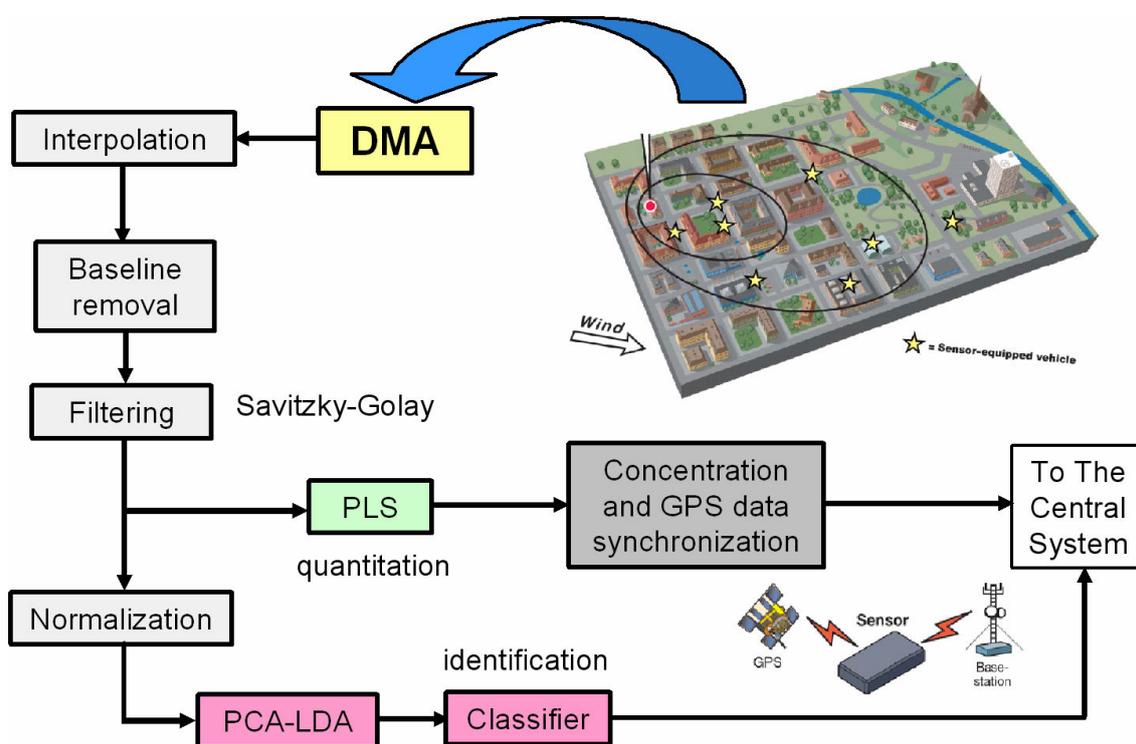
Figure 3.2. Block diagram of the DMA operating in real-time in a real environment (scenario described in chapter 1, 1.2.6).

3.5 DMA for explosive detection

3.5.1 Facilities

This study for explosive detection was carried out using the DMA manufactured by RAMEM in the "*Instituto Tecnológico La Marañosa* (ITM)" (Madrid, Spain) in different runs between May and November 2010.

The experiments are inspired in the improvised explosive detection scenario (section 1.2.2) and in the framework of the SEDUCE project (*Sistemas para la detección de explosivos en centros e infrastucturas públicas*).

3.5.2 Data description

Samples of different substances, mainly explosives, were measured in different runs in different days. Some of the measured substances were: acetone, DNT (Dinitrotoluene), TNT (Trinitrotoluene), NG (nitroglycerin), TATP (acetone peroxide), DMMP (Dimethyl MethylPhosphanate), PETN (pentrite). The DMA operated in positive and negative polarity.

3.5.3 Multivariate data processing for explosive identification

In order to study if different explosives can be discriminated, measurements of a certain substance obtained in one day are used to build training models, and measurements obtained in another day are used to test the model. By doing so, it is ensured that the influence of the substance is what is being studied, not cross-sensitivities which may pollute the shape of the spectra as temperature or humidity. When building calibration models, PCA is used as a first step for dimensionality reduction; the optimum number of components to build the model is based on the cumulative variance. Then LDA is applied using the sample scores returned by PCA and the class information (substance) of each sample, thus enhancing the discrimination between substances. Validation samples, the ones which must be predicted, are projected onto the PCA and LDA spaces built in the training step and a KNN classifier with K=1 is applied on the scores to assign a class (substance) to each validation sample. A confusion matrix can be obtained so as to compare the predicted classes vs. the real classes.

3.5.4 Experimental results

3.5.4.1 Experiment 1

This first experiment was conducted in May 2010 with the DMA operating in positive ion mode. The substances and the number of samples used in training and in validation are shown in Table 3.1. Training and validation samples were obtained in different days. Figure 3.3 show the mobility spectra obtained for the different substances.

Substance	# samples training	# samples validation
Acetone	3	3
TNT	8	4
Pentrite	9	4
Gunpowder	4	4

Table 3.1. Substances analyzed in experiment 1.
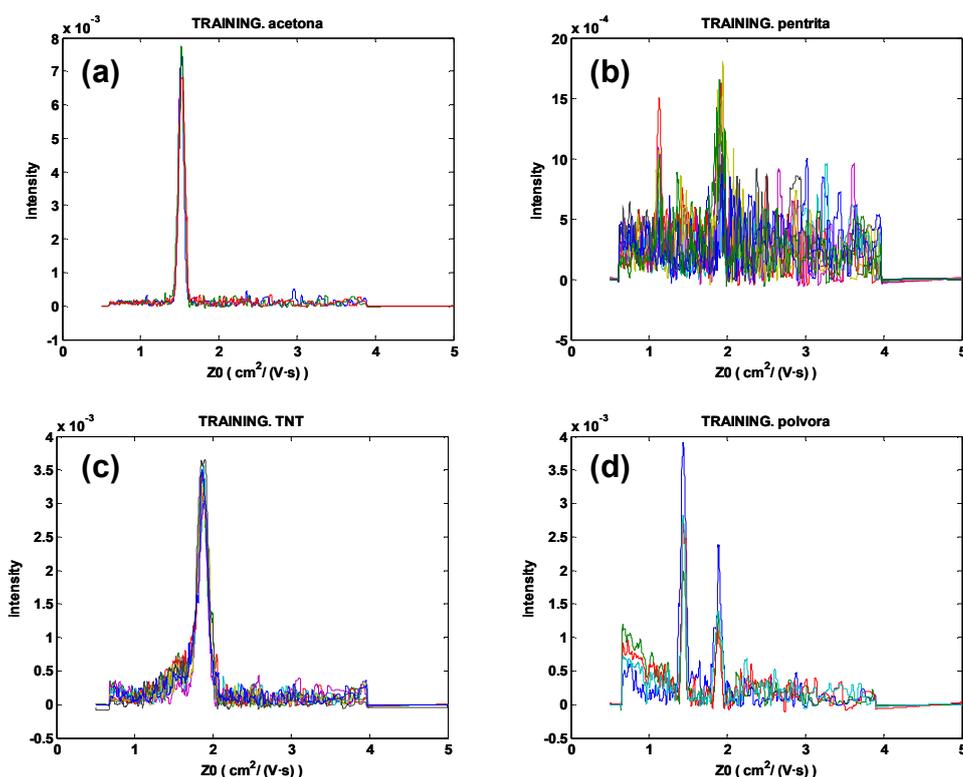
Figure 3.3. Mobility spectra of different substances (training data) obtained using the DMA in positive ion mode. (a) acetone. (b) pentrite. (c) TNT. (d) gunpowder.

It is seen in Figure 3.4(a) that samples of the different substances can be discriminated. This is verified in Table 3.2 where the predicted classes for each sample are compared with the real classes. It is observed that one validation sample of pentrite is classified as gunpowder and one validation sample of gunpowder is classified as pentrite. This is in agreement with Figure 3.4(b) where it is seen that pentrite and gunpowder samples are classified with low confidence values.
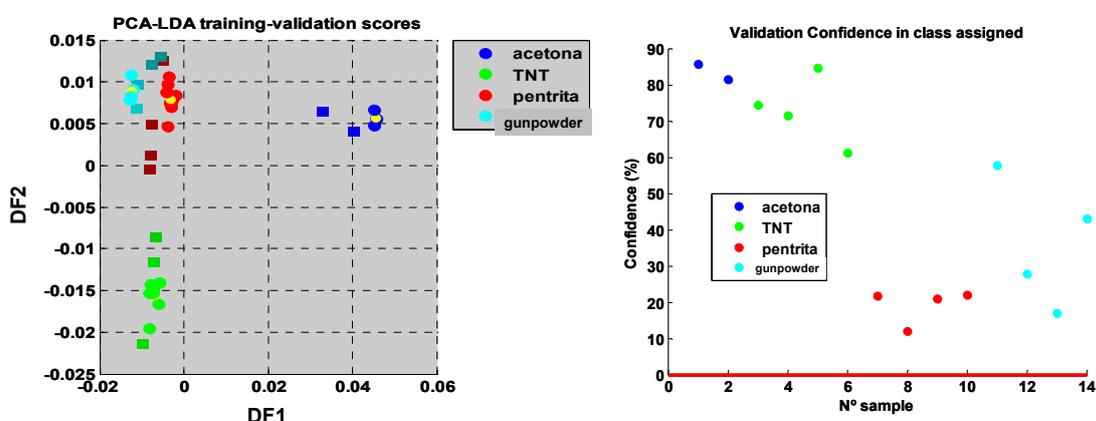
Figure 3.4. (a) Scores after PCA and LDA. Training samples ('o') and validation samples (squares). (b) Confidence in substance classification for each validation sample.

	CONFUSION MATRIX (VALIDATION)					
	Real class					
		Acetone	**TNT**	**Pentrite**	**Gunpowder**	**NONE**	**Outlier**
Acetone	2	0	0	0	0	0
TNT	0	4	0	0	0	0
Pentrite	0	0	3	1	0	0
Gunpowder	0	0	1	3	0	0
NONE	0	0	0	0	0	0
Outlier	0	0	0	0	0	0

(Predicted class labels down the left side)

Table 3.2. Confusion matrix for experiment 1.

The rate of overall classification was 85%.

3.5.4.2 Experiment 2

This second experiment was conducted between October and November 2010 with the DMA operating in positive ion mode. The substances and the number of samples used in training and in validation are shown in Table 3.3. Training and validation samples were obtained in different days.

Substance	# samples training	# samples validation
DMMP	10	4
TATP	4	2

Table 3.3. Substances analyzed in experiment 2.
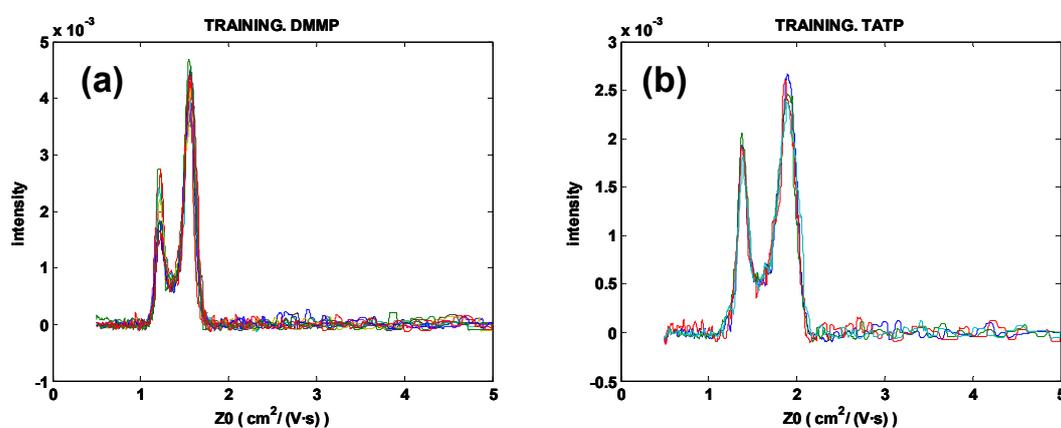
Figure 3.5. Mobility spectra of different substances (training data) obtained using the DMA in positive ion mode. (a) DMMP. (b) TATP.
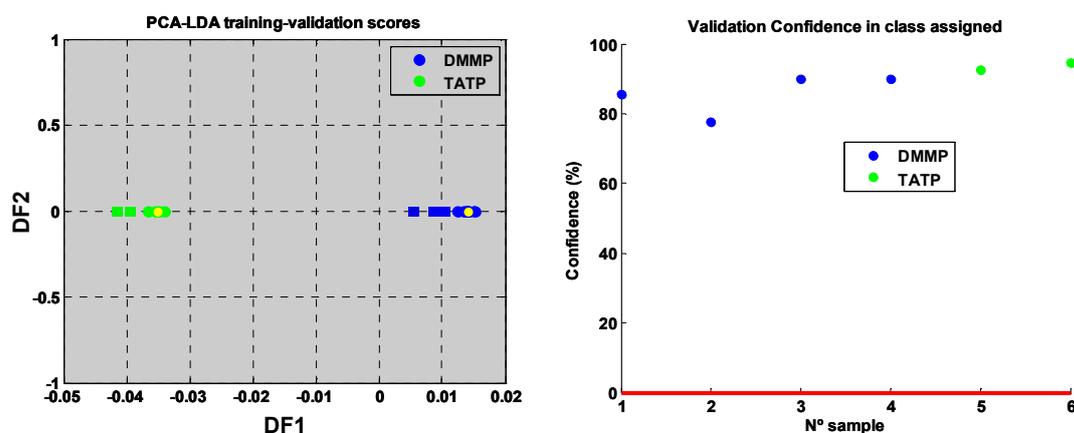
Figure 3.6. (a) Scores after PCA and LDA. Training samples ('o') and validation samples (squares). (b) Confidence in substance classification for each validation sample.

In this case it is seen that DMMP can be perfectly differentiate from TATP (Figure 3.6). Moreover, the confidence in the assigned class is pretty high for all validation samples. The confusion matrix is shown in Table 3.4. The rate of overall classification was 100%.

CONFUSION MATRIX (VALIDATION)					
		Real class			
		DMMP	TATP	NONE	outlier
Predicted class	DMMP	4	0	0	0
	TATP	0	2	0	0
	NONE	0	0	0	0
	outlier	0	0	0	0

Table 3.4. Confusion matrix for experiment 2.

3.5.4.3 Experiment 3

This third experiment was also conducted between October and November 2010 with the DMA operating in negative ion mode. The substances and the number of samples used in training and in validation are shown in Table 3.5. Training and validation samples were obtained in different days.

Substance	# samples training	# samples validation
DNT	9	6
NG	7	1
TATP	5	1
Trilite	6	5

Table 3.5. Substances analyzed in experiment 3.
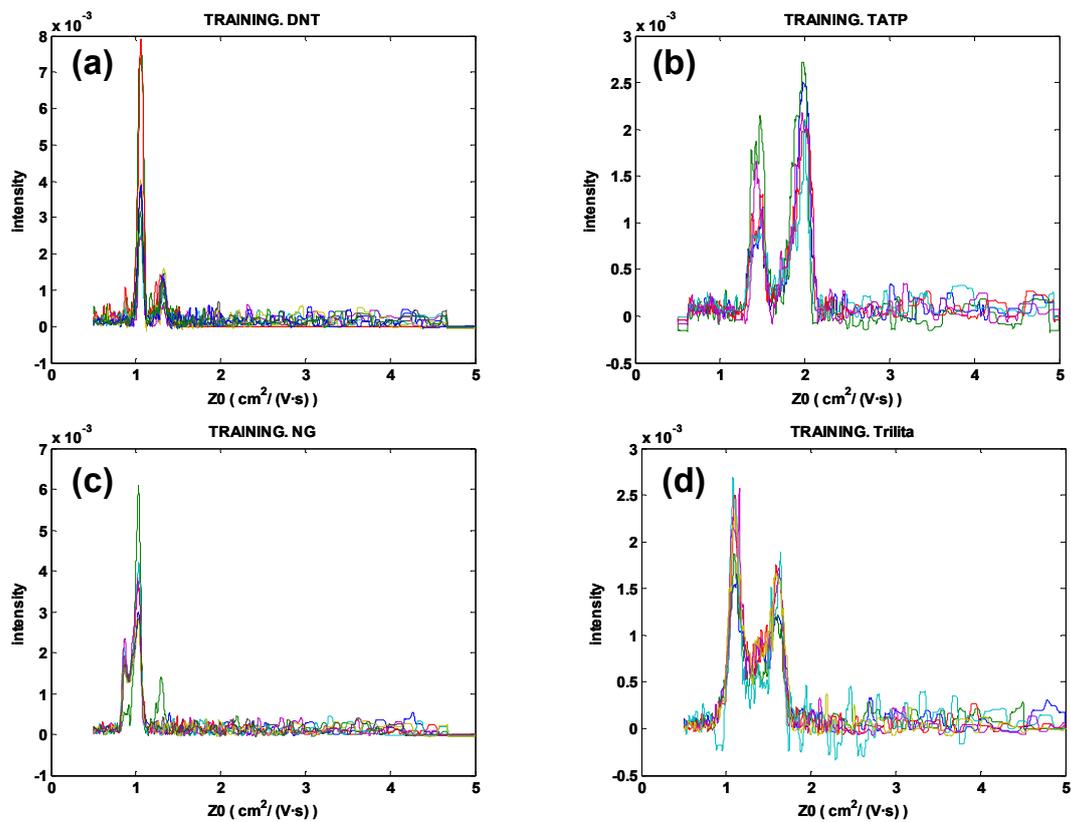
Figure 3.7. Mobility spectra of different substances (training data) obtained using the DMA in negative ion mode. (a) DNT. (b) TATP. (c) NG. (d) trilite.
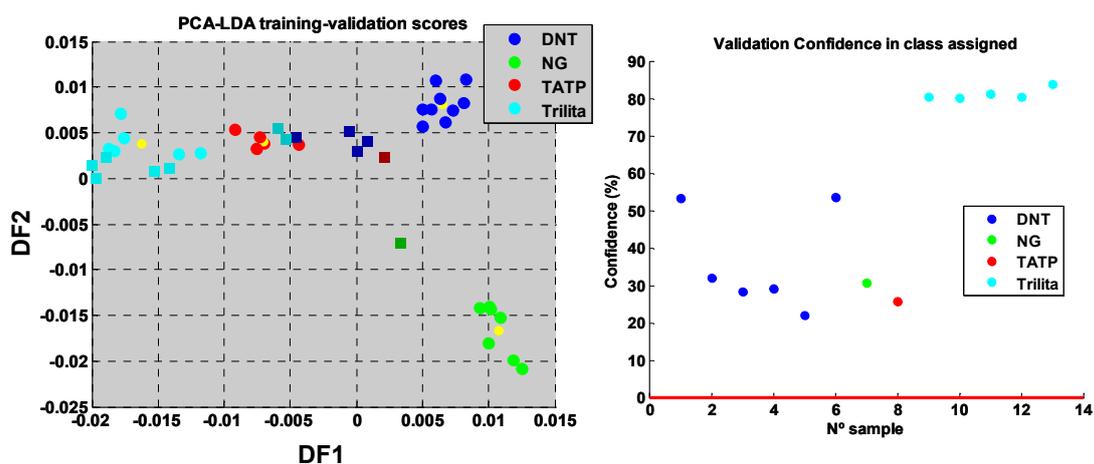
Figure 3.8. (a) Scores after PCA and LDA. Training samples ('o') and validation samples (squares). (b) Confidence in substance classification for each validation sample.

CONFUSION MATRIX (VALIDATION)							
		Real class					
		DNT	NG	TATP	Trilite	NONE	outlier
Predicted class	DNT	4	0	0	0	0	0
	NG	0	1	0	0	0	0
	TATP	0	0	1	0	0	0
	Trilite	2	0	0	5	0	0
	NONE	0	0	0	0	0	0
	outlier	0	0	0	0	0	0

Table 3.6. Confusion matrix for experiment 3.

The rate of overall classification was 85% with some DNT samples being classified as trilite (Table 3.6). This is in agreement with what is shown in Figure 3.8(b) where the confidence in substance classification of some DNT samples is low.

3.6 DMA for identification and quantitation of VOCs

3.6.1 Facilities and description of the experimental procedure

The measurements of the compounds for the present section were performed in one of the two chambers belonging to EUPHORE. EUPHORE (EUropean PHOto REactor) is one of the largest outdoor simulation chamber facilities in the world, used to study

atmospheric chemistry, belonging to the *"Centro de Estudios Ambientales del Mediterraneo* (CEAM)" in Valencia (Spain).

The tested volatile organic compounds (VOCs) were: acetone, benzene, toluene, ortho-xylene and para-xylene. P-xylene and toluene were purchased to Sigma Aldrich (Missouri, USA) with purities of 99% and 99.8%. O-xylene was purchased in Merck (New Jersey, USA) with purity of 98%. Benzene was bought in Scharlab (Barcelona, Spain) with purity 99% and acetone for HPLC to VWR with purity 99.8%.

Chemicals were injected into a PTFE chamber of clean air (200m^3) at increasing concentrations. The injection of the analytes was done using a small port where a glass flask was connected with a vial containing the analyte inside. A heater was used to evaporate the analyte and a gas flow was used to inject it into the chamber. Once inside the chamber, it was spread by the use of several fans. The chamber contents were continuously monitored with a calibrated FTIR analyzer (NICOLET). The FTIR signal for the analyte was followed until a constant signal was reached and at this point DMA measurements started. FTIR was also used to control the concentration of chemicals in the chamber. Concentration must be followed using calibrated techniques to consider the diminution that can be found by the sampling analyzers. FTIR cell is inside the chamber and is composed by White-type mirror system coupled to a FTIR spectrometer (NICOLET magna 550 MCTB/A detector). The total absorption path length is of 553.5m due to the reflection of the IR light inside the camera. FTIR data will be used as reference data to build PLS models for quantitation. The Limits of Detection (LOD) for the considered analytes are low enough for using them as reference data for the PLS model. LODs are 0.0060ppm for o-xylene, 0.0075ppm for benzene, 0.003ppm for acetone, 0.004ppm for toluene and 0.0019 por p-xylene. It must be remarked that the FTIR installed in the EUPHORE cameras is a quite complex device with two mirrors separated around 3m between them.

For DMA analysis, air was sampled from inside the chamber using a clean pump and PTFE pipes. 500ml·min^{-1} of air sampled from the chamber were injected inside the DMA without any sample preparation. After completion of the required experiments, in order to build each calibration curve, the air was humidified to check the spectral differences for each compound at higher concentrations and humidities. Once the chamber reached the desired humidity, generally 40%, the flushing process allowed measuring the calibration curve in the reverse mode, from higher concentrations to lower ones, and in this way, check the influence of humidity in the whole concentration range.

3.6.2 Data description

Acetone, benzene, o-xylene and toluene were measured on dry conditions (less than 15% relative humidity) at different levels of concentration. Acetone was measured at 14 increasing levels of concentration from 0.05 to 4ppm, with a total number of 116 DMA measurements. The range of concentrations measured for benzene was from 0.05 to 2.22ppm that included 7 different steps of concentrations. The total amount of measurements of benzene made with the DMA was 89. Up to 10 levels of concentration of o-xylene were measured from 0.05 to 3.05ppm, obtaining 69 DMA spectra. Finally, the studied range of concentrations of toluene was from 0.05 to 4.250ppm, made also in 11 levels and recording 124 measurements. Table 3.7 summarizes the measured concentrations.

Substance	Number scans	Conc. range (ppm)	Conc. levels	Number C_{levels} training	Number C_{levels} validation	Number scans training	Number scans validation
Acetone	116	0.05-4	14	8	6	66	50
Benzene	89	0.05-2.22	7	4	3	55	34
o-xylene	69	0.05-3.05	10	6	4	38	31
Toluene	124	0.05-4.25	11	6	5	66	58

Table 3.7. Measured substances at different concentrations on dry conditions (less than 15% of relative humidity).

Additionally, acetone, benzene and toluene were measured with relative humidity levels from 20 to 40%. p-xylene was measured with the same relative humidity levels instead of o-xylene. In this case the dataset comprises: 28 samples of acetone, 78 samples of benzene, 33 samples of p-xylene and 27 samples of toluene.

DMA datasets for each substance were pre-processed as explained in section 3.4.1.

3.6.3 Qualitative analysis

3.6.3.1 Methodology

Two pattern recognition models were built for qualitative analysis using the same training data used in quantitative analysis. Table 3.7 summarizes the number of concentration levels and the number of samples selected for training and for validation per each substance. Dimensionality reduction was based on either principal component analysis (PCA) or partial least squares (PLS-DA). A k-NN classifier was used to classify the samples in the validation subset after dimensionality reduction. In both cases, k (number of nearest neighbours) was taken as 3.

Leave-one-block-out (LOBO) methodology was used in order to establish the best number of principal components or latent variables that were used to build the dimensionality reduction models (see section 3.6.4.1). In this case, since a classifier was built, the total percentage of correct classification was taken as figure of merit. The number of components or latent variables which maximizes the percentage of classification is taken to build the calibration model.

Classification results are presented with the 95% confidence intervals. Additionally, results are presented in a confusion matrix that contains the classification rate per analyte, and the confusions between analytes. If the model predicts perfectly, the confusion matrix shows values with 100% in its diagonal. Different approaches have been proposed in the literature for the limit of detection (LOD) and quantification (LOQ) (Mocak et al. 1997; Vogelgesang & Hadrich 1998). Moreover, the used approaches may depend on the instrumentation (Tahboub, Zaater & Al-Talla 2005). A less common figure of merit is the limit of identification (LOI). In this work, we define the limit of identification (LOI) as the concentration above which the classification for a substance achieves 100% in the test set.

The effect of humidity

In order to see the effect of humidity in DMA samples, two analyses have been performed. In the first case, samples obtained under high humidity conditions are projected over the PLS-DA model built with samples at low humidity levels. In the second case, a specific PLS-DA model is built for samples at high humidity levels. In this case, training samples were selected randomly from the original set of samples with high humidity level. 70% of the samples were for the training dataset and 30% for validation. This cross-validation procedure is repeated 100 times, and the final percentages of classification are obtained averaging over these 100 trials.

3.6.3.2 Results

Table 3.8 contains the classification results of the two tested approaches (PCA and PLS-DA + KNN). The first approach that was based on PCA with a K-NN classifier had a total rate of good classification of 66% with 95% confidence intervals (57-72)%. This model was built using the first 2 principal components. The easiest analyte to detect was benzene with a rate of 97% with intervals (85-99)% of correct classification. The other three compounds were well detected in more than 40% of the samples. The LOI, as defined before (minimum concentration above which the analytes were perfectly assigned to their group), was estimated to be 0.7ppm, 0.9ppm, 1.5ppm and 2.0ppm for acetone, benzene, o-xylene and toluene, respectively.

The model based on PLS-DA using a K-NN classifier improved the total rate up to 77% with 95% confidence intervals (70-83)% for external validation samples using three latent variables. The improvement in the total rate of correct classification was done especially for benzene that is detected correctly in 94% of the test samples with confidence intervals (76-98)%. The identification limits were approximately equal to the PCA model using a KNN classifier for benzene and toluene, but it was decreased to 0.85ppm for o-xylene. PLS-DA-KNN approach provided a worse limit of identification than PCA-KNN approach in the unique case of acetone samples (1.5ppm).

	Substance	Classification intervals (%)	Limit of identification (ppm)
PCA-KNN model (2 PC, K=3) (Total rate = 66%)	Acetone	41-70	0.70
	Benzene	85-99	0.87
	o-xylene	25-61	1.55
	Toluene	52-77	2.00
PLS-KNN model (3LV, K=3) (Total rate = 77%)	Acetone	73-94	1.50
	Benzene	76-98	0.87
	o-xylene	33-70	0.85
	Toluene	63-86	2.00

Table 3.8. Summary of results for different qualitative models. Classification intervals are shown with 95% confidence value.

In general, PLS-DA-KNN provided better results in qualitative prediction of samples. The main reason is that building the PLS model, information about the substances is provided in an additional matrix. Whereas PLS tries to maximize the covariance between this extra matrix and the matrix of spectral responses, building a PCA model only the matrix of spectral responses is used. Figure 3.9 shows the PLS-DA scores plot. Scores projected over latent variable 1 and latent variable 2 are depicted. Training

samples are shown in black and validation samples in white. In Figure 3.9, all samples appear together for low concentrations since spectra did not show signal, whereas samples of the different substances can be identified increasing substance concentration. In order to analyze the results, Figure 3.9 should be used together with Table 3.9. Some samples of o-xylene and toluene are confused with acetone at low concentrations; moreover, some samples of o-xylene are confused with benzene at low concentrations.
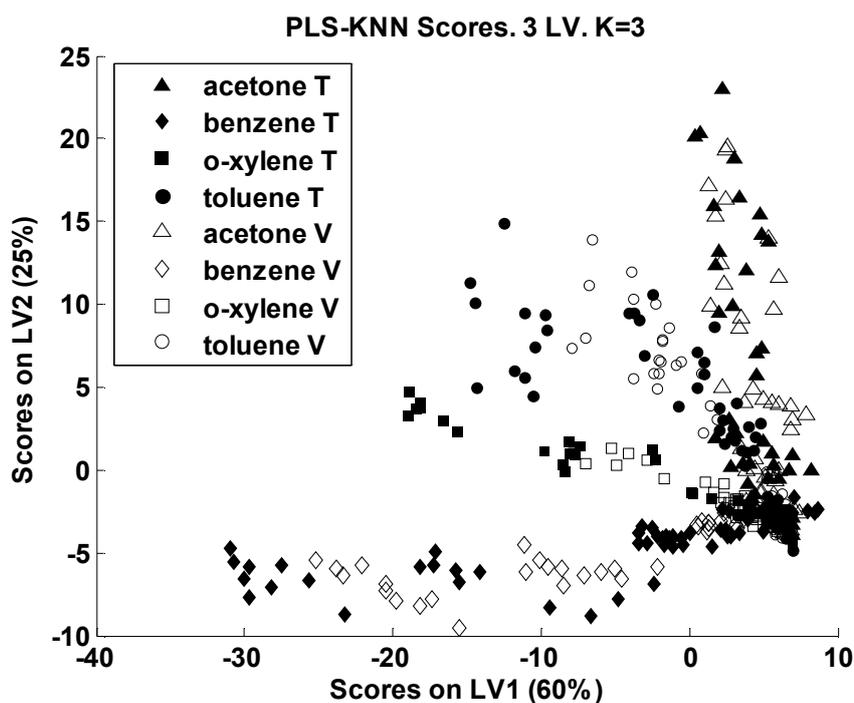
Figure 3.9. Score plot for PLS-DA classification model with 3 latent variables for all substances at low humidity level (less than 15%). Training samples are shown in black and validation samples in white.

Total rate = 77%	Expected acetone	Expected benzene	Expected o-xylene	Expected toluene
Obtained Acetone	86 (73-94)	3	29	20
Obtained benzene	2	94 (76-98)	16	2
Obtained o-xylene	4	3	52 (33-70)	2
Obtained toluene	8	0	3	76 (63-86)

Table 3.9. Confusion matrix for qualitative model PLS-DA-KNN with 3 latent variables and *K*=3 (Rate of classification in %) at low humidity level (less than 15%). 95% confidence intervals are shown in parentheses.

<u>The effect of humidity</u>

The visual inspection of the spectra with high levels of humidity (more than 15%) shows a peak at K_0=1.40cm^2/(V·s). This distortion in the spectra made the classification with humidity impossible. Samples measured at high humidity levels were projected over the PLS model built with samples at low humidity levels. All samples were classified as acetone due to the peak at K_0 =1.40cm^2/(V·s), therefore it is concluded that samples measured with high humidity levels cannot be classified with calibration models built with samples at low humidity levels. It is well known that humidity has a huge influence in mobility because ions at atmospheric pressure are likely to form hydration cluster, depending on their chemical nature. Hydration clusters of similar substances like benzene, toluene or xylene might drive to similar mobility values. Building a different model, enhancing the subtle differences for highly hydrated ions with a multivariate approach can be very useful when no membranes are used at the inlet of a mobility instrument. This is in agreement with other studies that have reported the increase of the detection limits with humidity in the sample using IMS (Vautz, Sielemann & Baumbach 2004).

When samples with high humidity level are used to build a new calibration model, results improve significantly. Figure 3.10 shows the score plot obtained from a PLS-DA model with 3 latent variables. Although Figure 3.10 shows a particular random selection it can be seen the averaged behaviour. On average (after 100 random selections) it is observed that some validation samples of p-xylene are classified as benzene (19%) and some samples of toluene are classified as p-xylene (15%). The averaged classification rate (Table 3.10) after 100 random selections of each one of the substances were: 98% for acetone, 93% for benzene, 71% p-xylene and 82% toluene, with 95% confidence intervals: (97-99)% for acetone, (92-94)% benzene, (68-74)% p-xylene and (79-84)% toluene. The averaged classification rate was 87% with 95% confidence intervals (86-88)%.

Total rate = 87%	Expected acetone	Expected benzene	Expected p-xylene	Expected toluene
Obtained Acetone	98 (97-99)	0	2	1
Obtained benzene	0	93 (92-94)	19	2
Obtained p-xylene	1	6	71 (68-74)	15
Obtained toluene	1	1	8	82 (79-84)

Table 3.10. Confusion matrix for qualitative model PLS-DA-KNN with 3 latent variables and K=3 (Rate of classification in %) at high humidity level (more than 15%). 95% confidence intervals are shown in parentheses.
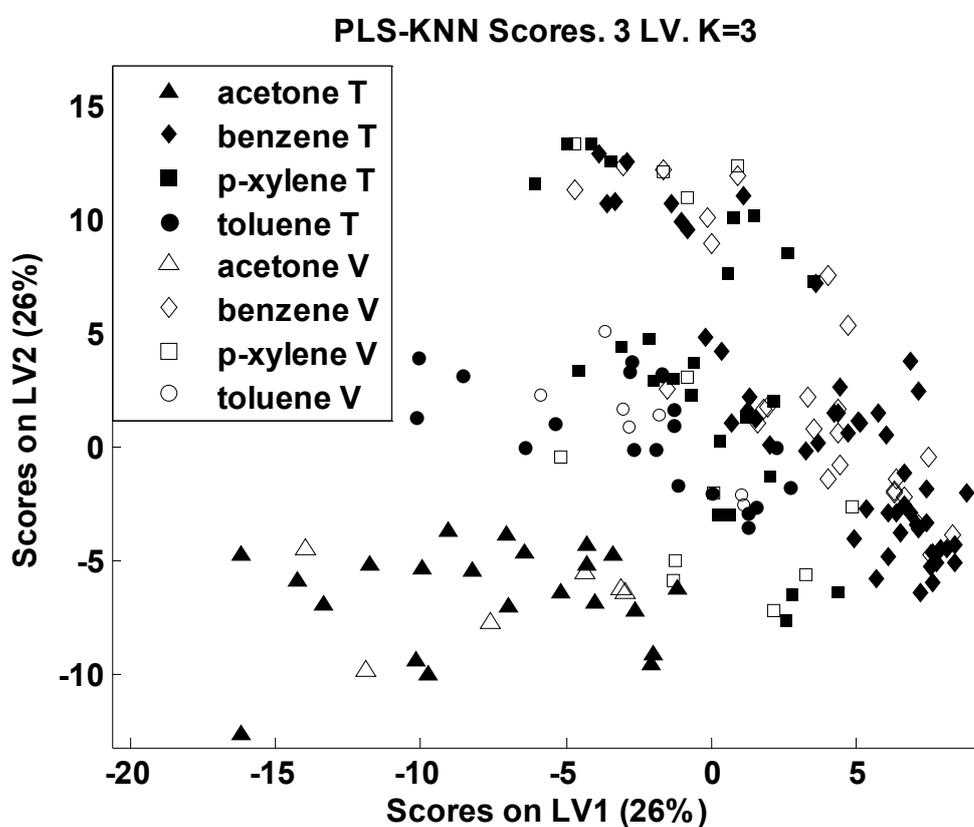
Figure 3.10. Score plot for PLS-DA classification model using 3 latent variables. Samples with high humidity levels (more than 15%). One random selection of samples (70% for training and 30% for validation). Training samples are shown in black and validation samples in white.

3.6.4 Quantitative analysis

3.6.4.1 Methodology

Three different approaches were used for obtaining quantitative models of acetone, benzene, o-xylene and toluene. All of them were based on Partial Least Squares (PLS) method (Geladi & Kowalski 1986). Reference concentrations were taken from the FTIR data.

The first approach consisted of building a single PLS model for each compound, without taking into account the contribution of other compounds.

The second one was also a single PLS model for each compound but including the rest of compounds as interferences. In this case, the spectra of interfering compounds were included in the model assuming that concentration of the substance of interest was zero in these interfering samples.

The last quantitative model tested was a unique PLS model including all compounds.

The set of samples were divided into two subsets, one training subset for building the calibration model and another validation subset for evaluating the predictive ability of the PLS models.

In the PLS models, the best number of latent variables was selected using the leave-one-block-out (LOBO) methodology on the training subset of samples, that is for a given substance, the whole set of spectra corresponding to a certain concentration is all taken out to be validated and the remaining set of samples is used to build the calibration model. The set of spectra corresponding to the first and last measured concentrations are always used to build the calibration model avoiding the possibility to predict concentrations out of the calibration range.

The set of scans to be validated is used to calculate and minimize the root mean squared error of validation (RMSEV):

(Eq. 3.4)
$$RMSEV = \sqrt{\frac{\sum_{i=1}^{N}\left(C_i - \widehat{C}_i\right)^2}{N}}$$

where C_i is the concentration of the sample, $\widehat{C}_i$ is the predicted concentration of the sample by the model, and N is the number of validated samples. The optimum number of components to build the calibration model is the one which minimizes the RMSEV.

In order to compare quantitative results among the different substances, the RMSEV can be expressed as a percentage of the maximum substance concentration.

3.6.4.2 Results

Three different approaches for quantification of acetone, benzene, o-xylene and toluene were obtained with the same data.

The first one corresponds to four different PLS models; each one was built without taking into account the other three compounds. As it can be observed in Table 3.11, the regression of the obtained vs. expected concentrations for the test samples shows the ideal behaviour, with slopes close to one, intercept close to zero and high correlations. Moreover, the RMSEV values are quite low: 0.15ppm (3.8%), 0.18ppm (8.6%), 0.14ppm (4.6%) and 0.33ppm (7.8%) for acetone, benzene, o-xylene and toluene, respectively; which means that the quantification errors are low on average. However, these models built using only one substance do not incorporate information about the interference of the other compounds and thus they are not useful in a real scenario.

	Substance	Slope	Intercept (ppm)	R	RMSEV(% max conc)	#LV opt
Individual PLS model	Acetone	1.05 ± 0.02	-0.11 ± 0.03	0.99	3.8	4
	Benzene	0.91 ± 0.04	0.10 ± 0.05	0.96	8.6	2
	o-xylene	0.98 ± 0.06	-0.04 ± 0.04	0.95	4.6	3
	Toluene	0.89 ± 0.03	-0.10 ± 0.05	0.96	7.8	2
Individual PLS model with other substances at 0 concentration	Acetone	0.92 ± 0.02	-0.01 ± 0.02	0.97	5.3	6
	Benzene	0.94 ± 0.02	-0.01 ± 0.01	0.96	6.3	9
	o-xylene	0.83 ± 0.04	0.02 ± 0.01	0.86	4.9	8
	Toluene	0.90 ± 0.02	0.02 ± 0.02	0.95	5.2	12
Global PLS model	Acetone	0.92 ± 0.02	-0.004 ± 0.02	0.96	5.8	7
	Benzene	0.96 ± 0.02	-0.01 ± 0.01	0.96	6.3	7
	o-xylene	0.81 ± 0.04	0.02 ± 0.01	0.86	5.2	7
	Toluene	0.91 ± 0.02	-0.003 ± 0.02	0.96	4.9	7

Table 3.11. Summary of results for different quantitative models. RMSEV is presented as a percentage of the maximum substance concentration. Errors in slope and intercept are given within one standard deviation.

The second studied approach is based also in four different PLS models, each one for each substance but in this case the samples of the other three substances were added to the model as samples with zero concentration for the studied substance, which means that they were considered as interferences. The optimum number of latent variables to build the calibration models was found to be higher in the second approach. The reason is that more components are needed in order to explain nonlinearities due to the presence of interfering substances. This set of individual PLS models performed well for quantifications of acetone, benzene and toluene (see their slopes in Table 3.11), and only slightly well for o-xylene, with a slope of 0.83. The RMSEV values of this set of models, reported in Table 3.11, are quite similar to those obtained using the first approach. Although this set of models worked less successfully for each single compound in absence of interferences, the study of cross-sensitivity to interference substances shows that can be used in more realistic situations because the addition of interferences in the training step improved their selectivity.

The last tested quantitative approach was based on unique PLS model that included the information of the four compounds. The optimum number of latent variables was found to be 7. The comparison between obtained and expected concentrations (see also Table 3.11 and Figure 3.11) has similar slope values, intercepts and correlations as the second approach and the RMSEV are also similar. The unique PLS model also takes into account interferences and shows good cross-sensitivity behaviour. The main reason for this is that although spectra of the different substances look like very similar among them, using a high number of latent variables, the model is able to capture unique information for each substance thus providing good predictive results.
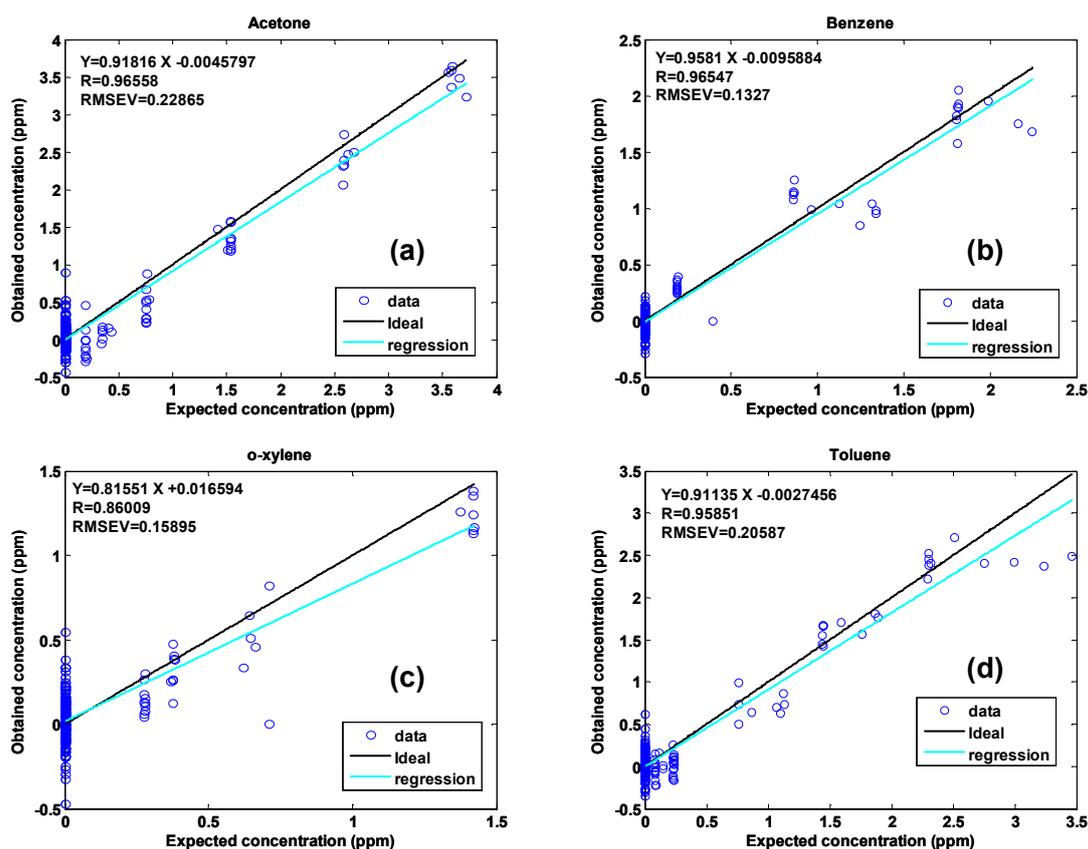
Figure 3.11. Quantitative results for the unique PLS model. (A) Acetone; (B) Benzene; (C) o-Xylene; (D) Toluene. (Concentrations and RMSEV in ppm).

3.7 Conclusions

In this chapter, the first measurements with a stand-lone DMA instrument and the analysis of the spectra with multivariate processing techniques are presented. In section 3.5 explosives are correctly identified and in section 3.6 different VOCs are identified and quantified.

Although the presented experimental results in section 3.5 show that the system (DMA + multivariate data processing) is able to correctly discriminate among different explosives, the low number of samples makes that this study can only be considered as preliminary. New runs of experiments should be carried out in the future in order to establish the DMA by RAMEM as effective tool for explosive detection.

In section 3.6, four VOCs (acetone, benzene, o-xylene and toluene) has been detected, identified and quantified with the DMA instrument with good performance. Limits of identification (LOI) of 0.7ppm for acetone, 0.9ppm for benzene, 1.5ppm for o-xylene and 2.0ppm for toluene were found. The PLS models have been able to quantify the

four analytes. The RMSEV for acetone, benzene, toluene and o-xylene were in the range from 0.1 to 0.3ppm in validation. Identification and classification of the compounds has been done based on PLS-DA with a K-NN classifier built on the latent variable space. The rate for global classification was 77% with 95% confidence intervals of (70-83)%. However, most of the confusions are due to the concentrations that are below the limit of identification.

Current results also show the great influence of the humidity. Next DMA prototype will include conditioning of the sheath gas, with heating elements and a filter to test several molecular sieves. This will reduce the great influence of the humidity in the resulting spectra, and more reproducible results are expected.

Furthermore, the general procedure to predict (identification and quantitation) unknown samples in a real environment in real-time is described. First, training models are built using data obtained from the DMA under controlled conditions in the laboratory; second, these models are used in the real environment to predict the new measurements in real-time. Although this procedure has been explained for the DMA case, since it is general, it can be applied to any instrumentation providing multi-dimensional data (gas sensor arrays, IMS or spectroscopy in general).

We would like to remark that a full suite of tools for DMA spectra processing, including pre-processing, dimensionality reduction, classifiers and regressors has been developed in Mathscript (LabView) from scratch and integrated on the on-board LabView programme that currently is offered with the instrument by the producer.

3.8 References

Boger, Z & Karpas, Z 1994, 'Use of neural networks for quantitative measurements in ion mobility spectrometry (IMS)', *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 3, pp. 576-580.

Bota, GM & Harrington, PB 2006, 'Direct detection of trimethylamine in meat food products using ion mobility spectrometry', *Talanta*, vol. 68, no. 3, pp. 629-635.

EPA, *http://www.epa.gov/*, US Environmental Protection Agency.

Fraga, CG, Kerr, DR & Atkinson, DA 2009, 'Improved quantitative analysis of ion mobility spectrometry by chemometric multivariate calibration', *Analyst*, vol. 134, no. 11, pp. 2329-2337.

Gan, F, Ruan, G & Mo, J 2006, 'Baseline correction by improved iterative polynomial fitting with automatic threshold', *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 59-65.

Geladi, P & Kowalski, BR 1986, 'Partial least-squares regression: a tutorial', *Analytica Chimica Acta*, vol. 185, no. 0, pp. 1-17.

Karpas, Z, Guamán, AV, Calvo, D, Pardo, A & Marco, S 2012, 'The potential of ion mobility spectrometry (IMS) for detection of 2,4,6-trichloroanisole (2,4,6-TCA) in wine', *Talanta*, vol. 93, no. 0, pp. 200-205.

Mocak, J, Bond, AM, Mitchell, S & Scollary, G 1997, 'A statistical overview of standard (IUPAC and ACS) and new procedures for determining the limits of detection and quantification: Application to voltammetric and stripping techniques (technical report)', *Pure and Applied Chemistry*, vol. 69, no. 2, pp. 297-328.

NIST, *http://www.nist.gov/index.html*, National Institute of Standards and Technology.

Ochoa, ML & Harrington, PB 2005, 'Chemometric studies for the characterization and differentiation of microorganisms using in situ derivatization and thermal desorption ion mobility spectrometry', *Analytical Chemistry*, vol. 77, no. 3, pp. 854-863.

Pomareda, V, Calvo, D, Pardo, A & Marco, S 2010, 'Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra', *Chemometrics and Intelligent Laboratory Systems*, vol. 104, no. 2, pp. 318-332.

Pomareda, V, Guamán, AV, Mohammadnejad, M, Calvo, D, Pardo, A & Marco, S 2012, 'Multivariate curve resolution of nonlinear ion mobility spectra followed by multivariate nonlinear calibration for quantitative prediction', *Chemometrics and Intelligent Laboratory Systems*, vol. 118, no. 0, pp. 219-229.

Prasad, S, Pierce, KM, Schmidt, H, Rao, JV, Guth, R, Synovec, RE, Smith, GB & Eiceman, GA 2008, 'Constituents with independence from growth temperature for bacteria using pyrolysis-gas chromatography/differential mobility spectrometry with analysis of variance and principal component analysis', *Analyst*, vol. 133, no. 6, pp. 760-767.

RAMEM, *http://www.ioner.eu/*, RAMEM S.A.

Rauch, PJ, Harrington, PD & Davis, DM 1998, 'Chemometric resolution of mixture components by cleardown rates', *Analytical Chemistry*, vol. 70, no. 4, pp. 716-723.

Reese, ES & Harrington, PD 1999, 'The analysis of methamphetamine hydrochloride by thermal desorption ion mobility spectrometry and SIMPLISMA', *Journal of Forensic Sciences*, vol. 44, no. 1, pp. 68-76.

Savitzky, A & Golay, MJE 1964, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', *Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639.

Tahboub, YR, Zaater, MF & Al-Talla, ZA 2005, 'Determination of the limits of identification and quantitation of selected organochlorine and organophosphorous pesticide residues in surface water by full-scan gas chromatography/mass spectrometry', *Journal of Chromatography A*, vol. 1098, no. 1-2, pp. 150-155.

Vautz, W, Sielemann, S & Baumbach, JI 2004, 'Determination of terpenes in humid ambient air using ultraviolet ion mobility spectrometry', *Analytica Chimica Acta*, vol. 513, no. 2, pp. 393-399.

Vidal-de-Miguel, G, Macía, M & Cuevas, J 2012, 'Transversal Modulation Ion Mobility Spectrometry (TM-IMS), A New Mobility Filter Overcoming Turbulence Related Limitations', *Analytical Chemistry*, vol. 84, no. 18, pp. 7831-7837.

Vogelgesang, J & Hadrich, J 1998, 'Limits of detection, identification and determination: a statistical approach for practitioners', *Accreditation and Quality Assurance*, vol. 3, no. 6, pp. 242-255.

Zamora, D & Blanco, M 2012, 'Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration', *Analytica Chimica Acta*, vol. 726, no. 0, pp. 50-56.

Zheng, P, Harrington, PdB & Davis, DM 1996, 'Quantitative analysis of volatile organic compounds using ion mobility spectrometry and cascade correlation neural networks', *Chemometrics and Intelligent Laboratory Systems*, vol. 33, pp. 121-132.

Chapter 4

Sparse MCR-ALS using L1 regularization and Gaussian peak shape models

4.1 Introduction

As it was described in chapter 1, Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) aims to recover the evolution of the source signals (in our case, concentration profiles) and the mixing matrix (spectral features) without any prior supervised calibration step. Additionally, it is well known that imposing additional knowledge about the methods, or the processes, by using constraints can lead to better solutions and easier interpretation of the results since rotational ambiguities (Jaumot & Tauler 2010) are minimized and the space of possible solutions is reduced. In this sense, constraining the solution by imposing several constraints is a standard practice in MCR (Juan et al. 1997). For instance, some soft-constraints (Tauler 1995; Gemperline & Cash 2003; Bro & Sidiropoulos 1998) are: 1) number of pure components expected to be found in the mixture, 2) Non-negativity, 3) Unimodality, 4) Selectivity and 5) Closure. Nevertheless, when using IMS, in a real scenario where interfering substances are expected to be present, it is almost impossible to known in advance if a spectrum is unimodal or which regions are selective to one or more specific ions. Therefore, in the experiments presented in this chapter, only the constraints 1 and 2 were applied so as to force the solution as less as possible.

While, in its basic form, MCR is a soft-modelling technique (not underlying model is imposed), several authors have proposed hard-modelling versions of MCR where physical-chemical models characterizing the underlying process are imposed in the solution (Juan et al. 2000). When using these models, additional knowledge is incorporated, therefore the solution is even more constrained (hard constraint) and rotational ambiguities are minimized as well.

For some spectroscopic measurements, spectra are characterized by the presence of a series of peaks. In some instruments, approximate models of the peak shape are known (e.g. Gaussian peaks) (Felinger 1998). However, in a blind source separation (BSS) scenario, neither the number of peaks nor their position is known in advance. This point seriously hinders the application of peak shape models within the alternating least square solution loop of MCR.

We propose to model the source spectra as a dense superposition of Gaussian peaks and then rely on L1-norm regularization to obtain a sparse solution in the model, thus obtaining automatically the proper number of peaks and their location without imposing a priori either the location or the number. To do so, we have introduced the LASSO (Least Absolute Shrinkage and Selection Operator) technique within the MCR-ALS loop. We call the new algorithm MCR-LASSO. LASSO was proposed by Tibshirani (Tibshirani 1996) and it is known as *basis pursuit* (Chen, Donoho & Saunders 1998) or *compressed sensing* (Donoho 2006) in the field of signal processing. To test this concept we have applied basic MCR-ALS and MCR-LASSO to both synthetic and real signals. Real signals correspond to Ion Mobility Spectra (IMS) recorded using a baggage scanner prototype which mimics an airport security checkpoint (section 4.3.4).

4.2 New proposed BSS technique: MCR-LASSO

The main motivation to propose a new BSS technique is to improve the quality of the estimation using the less possible number of constraints. The shape of IMS spectra is very dependent on the combination of substances present in a sample. Moreover, in real scenarios, where interfering chemicals are present, the spectra could become even more complicated, thus it is unknown if some constraints can be applied to IMS spectra, for instance: unimodality, selectivity or local chemical rank.

The proposed MCR-LASSO technique is based on MCR. A flexible mathematical model for the spectrum shape is introduced in the algorithm and the complexity of this model is controlled by a regularization parameter. Spangler showed that, for IMS, the shape for a migrating ion cloud arriving at the collector could be considered Gaussian as a first approximation. If there are losses due to ion conversion, recombination or transverse diffusion, the peak gives up being Gaussian and gets wider (Spangler 2002). In our model, spectrum S_j is modeled as a linear superposition of Gaussians of variable width (regressors):

(Eq. 4.1)
$$S_j(t) = \sum_{i=1}^{N} a_{ij} \varphi_i(t)$$

where:

(Eq. 4.2)
$$\varphi_i(t) = \exp\left[\frac{-(t - t_i)}{2 \cdot \sigma_i^2}\right]$$

forms the basis of regressors which is used for the linear regression problem; t is the drift time, t_i is the centre of each Gaussian in the *x*-axis, N is the number of Gaussians (eigenvectors) to reconstruct a spectrum, and

(Eq. 4.3)
$$\sigma_i = \frac{t_i}{R_p \cdot 2 \cdot \sqrt{2 \cdot \ln 2}}$$

is the standard deviation of each one of these Gaussians located on each drift time point. R_p is the resolving power of the IMS defined in Eq. 1.3 and in (Rokushika et al. 1985; Siems et al. 1994; Spangler 2002). N is equal to the number of drift time points per spectrum.

This dense linear combination of Gaussians has the capability to model a large diversity of peaks. The only constraint being that they cannot be narrower than σ_i.

It is very important to understand that the only fitting parameters are the set of weights $\{a_i\}$ in (Eq. 4.1). In such a way, the model becomes linear in the parameters. We rely on this dense Gaussian model to fit wider non-Gaussian peaks. Please note, that in our approach N is very large, or in other words, the sampling time $\Delta=t_{i+1}-t_i$, is much smaller than the peak FWHM (Full Width at Half Maximum). While the fitting parameters $\{a_i\}$ can be estimated by ordinary least squares, this estimation is ill conditioned when the correlation among the regressors is large (Eldén 1977; Golub 1965). The condition number with respect to inversion gives us estimation on the sensitivity of the solution of a system of linear equations to errors in the data. The condition number is defined as the ratio of the largest singular value of the matrix of regressors (Eq. 4.2) to the smallest. Large condition numbers indicate a nearly singular matrix. In our spectrum model, each Gaussian is highly correlated with a number of neighbouring Gaussians and the condition number is around $3 \cdot 10^{19}$. This fact can lead to singularities and instabilities producing a poor estimation for the a_{ij} coefficients, which will exhibit high variance. A large positive coefficient in one regressor can be cancelled by a similarly large negative coefficient in a correlated neighbour. The problem can be regularized by imposing a constraint on the coefficients.

Regularization to MCR-ALS solution has been previously considered by Wang (Wang et al. 2003) using Ridge Regression which does a proportional shrinkage to all the coefficients. However, in order to have a sparse solution minimizing the number of Gaussians in the model, we propose to use LASSO (Tibshirani 1996; Hastie, Tibshirani & Friedman 2009). LASSO is an iterative least squares solver using an L1-norm penalty. The LASSO shrinks some fitting coefficients and sets many others to zero. It is accepted that LASSO produces more parsimonious models providing better prediction accuracy and more interpretable models (Hastie, Tibshirani & Friedman 2009).

We propose to introduce LASSO within Multivariate Curve Resolution in the estimation of pure spectra in each ALS iterative step. Unlike the ridge regression, which uses a L2-norm penalty in the coefficients ($\sum a_j^2$), the LASSO does not have a closed form

107

because it uses a L1-norm ($\sum |a_j|$); however an estimation can be derived from a linear approximation in such a way that the vector of weights $\boldsymbol{a}_j$ for the j-th spectrum can be written as the ridge regression estimator ($|a_j| \cong \dfrac{a_j^2}{|a_j|}$) and the solution can be found by iteratively computing the ridge regression (Fan & Li 2001):

$$(Eq.\ 4.4) \qquad \vec{a}_{j,new} = \left[\varphi^T \varphi + \lambda \cdot diag\left(\left| \vec{a}_{j,old} \right| \right)^{-1} \right]^{-1} \varphi^T S_j$$

where φ is the regression matrix of gaussians (Eq. 4.2) at each time; λ is the tuning or regularization parameter which, in general, is adjusted by cross-validation (Hastie, Tibshirani & Friedman 2009). The tuning parameter $\lambda \geq 0$ controls the amount of shrinkage on the coefficients. The larger the value of λ, the greater the amount of shrinkage. Moreover, λ regularizes the estimation of the a_{ij} coefficients adding a positive constant to the diagonal of $\varphi^T \varphi$ before inversion (Eq. 4.4). This makes the problem non-singular, even if $\varphi^T \varphi$ is not of full rank, which is our case. For a description about linear methods for regression and specifically shrinkage methods, the reader is referred to (Tibshirani 1996; Hastie, Tibshirani & Friedman 2009).

4.3 Materials and Methods

4.3.1 MCR-LASSO algorithm

MCR-LASSO algorithm starts filtering the data matrix with PCA (using the same number of pure components which will be used later in the main loop) and using the first estimations for spectra from SIMPLISMA (Windig et al. 2005). Then C is obtained by means of fast non-negative least squares (FNNLS) (Bro & Jong 1997) from (Eq. 4.5), and S is estimated using C and FNNLS (which impose non-negativity in a least squares sense).

$$(Eq.\ 4.5) \qquad\qquad D = C \cdot S^T + E$$

S is normalized to unit area before being used by LASSO, which, given a penalty parameter λ, estimates the best a_{ij} coefficients for each spectrum using (Eq. 4.4) iteratively. At this point, the most of the a_{ij} coefficients are zero, generating a sparse solution producing a small and stable subset of Gaussians even in the presence of noise. Using these a_{ij} coefficients, spectra are reconstructed and consequently filtered,

removing most of the high frequency noise, and normalized to unit area. The algorithm enters in an iterative procedure recalculating *C*, *S* and a_{ij} and goes on until it converges (the relative difference between root mean squared errors (RMSE) for successive iterations is small enough) or the maximum number of iterations is achieved. The RMSE is defined as:

(Eq. 4.6)
$$RMSE = \sqrt{\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(D_{m,n} - \hat{D}_{m,n}\right)^2}{M \cdot N}}$$

where *D* is the original data matrix filtered with PCA and $\hat{D}$ is the reconstructed data matrix using *C* and *S* for the current iteration (Eq. 4.5).

The block diagram of the MCR-LASSO algorithm is given in Figure 4.1, which is the same as that of MCR-ALS but for the additional LASSO step. Thereby, intuitively, when *λ* tends to zero, the MCR-LASSO solution tends to the MCR-ALS solution with slight differences due to the rigidity of the Gaussian model.
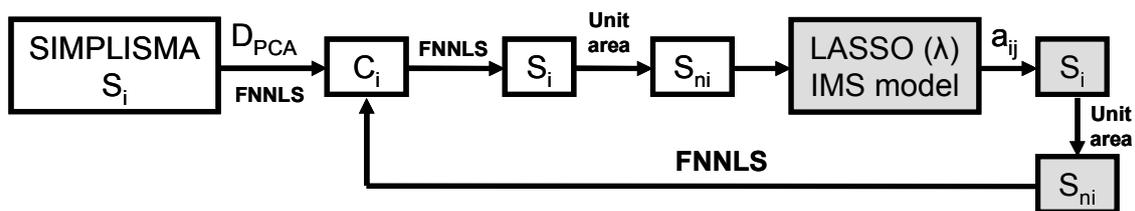
Figure 4.1. Block diagram of MCR-LASSO algorithm. The additional blocks added with regard to MCR-ALS algorithm are shown shaded.

4.3.2 Non-Gaussian peak fitted from a Gaussian model using LASSO

First, we want to prove that a non-Gaussian peak with a long tail can be modelled as a linear superposition of Gaussian peaks. Secondly, we want to prove that LASSO is more suitable than OLS (Ordinary Least Squares) for fitting purposes (a_{ij} determination) in the presence of noise. In order to demonstrate the previous hypotheses, a non-Gaussian peak is synthetically generated. We use the Breit-Wigner function, also known as Cauchy or Lorentz distribution, previously applied to model IMS signals (Vogtland & Baumbach 2009). A peak is composed of a left Gaussian part and a right Breit-Wigner part which models the peak's tail. A measurement of asymmetry is the ratio between the FWHM and the FWHM for a Gaussian of the model at the same position. For the current non-Gaussian peak this ratio is 1.51 (a value of 1 would indicate that the peak is perfectly Gaussian). LASSO and OLS are tested under noisy

conditions. Additive normally distributed noise is added to the spectrum to reach a signal-to-noise ratio (SNR) of 10dB. The SNR is defined as:

(Eq. 4.7)
$$SNR(dB) = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

where P_{signal} and P_{noise} are the mean power of the signal and the noise, respectively.

In the case of LASSO, since, in this example, the solution is known and so as to demonstrate the hypothesis, the tuning parameter λ is determined from a sweep. Given λ, the RMSE (Eq. 4.6) is evaluated between the original asymmetrical peak and the LASSO fit. The λ which provides the minimum RMSE is selected.
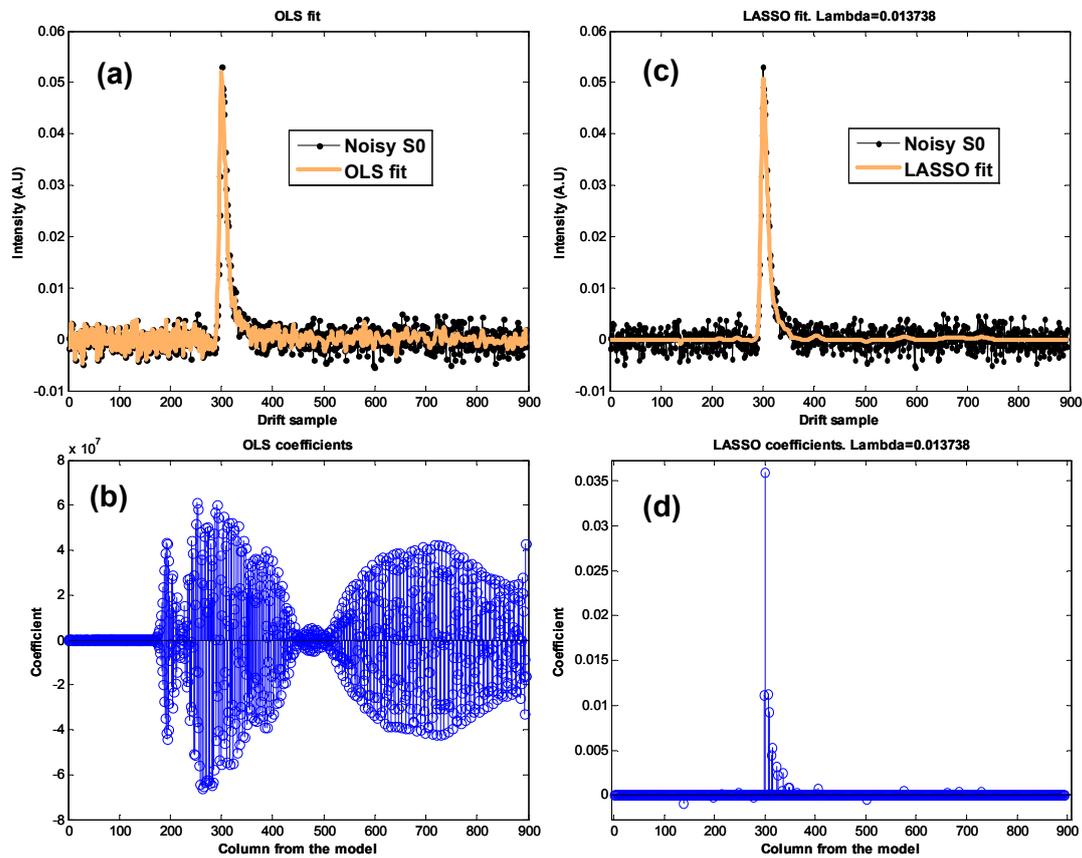
Figure 4.2. Comparison between OLS and LASSO fit for a noisy (SNR=10dB) asymmetrical peak (half width at half height left part w_g=4.7 and half width at half height right part w_{bw}=9.5). (a) OLS fit compared to original noisy peak. (b) OLS coefficients. (c) LASSO fit (λ=0.014) compared to original noisy peak. (d) LASSO shrunk coefficients, giving a small and stable subset of Gaussians to reconstruct the original peak.

Figure 4.2 shows the fitting results for OLS (Figure 4.2(a)) and LASSO (Figure 4.2(c)) using the dense Gaussian model as basis of regressors. The long tail is fitted in both approaches. Even though the noisy conditions and the high asymmetry in the original peak, the LASSO solution is able to fit the peak giving a RMSE of $3 \cdot 10^{-4}$ and only 3% of the coefficients are different from 0. The OLS solution is fitting the noise giving a RMSE of $1 \cdot 10^{-3}$ and all the coefficients are different from 0 (100%). Moreover, the coefficients are much more stable and regularized in the LASSO solution (Figure 4.2(d)) with much smaller values than in the OLS solution (Figure 4.2(b)). In the OLS solution, there are a great number of positive and negative coefficients trying to compensate each other; this shows the instability of the solution if there is a great correlation among regressors, which leads to a poor estimation of the coefficients. Unlike OLS, the LASSO provides an optimal, small and stable subset of Gaussians to reconstruct the original asymmetrical peak in the presence of noise.

4.3.3 Synthetic dataset

A collection of datasets (training, validation and test) of synthetic spectra were generated to test the algorithms under noisy conditions (Gaussian additive noise to reach SNR=10dB (Eq. 4.7) and highly overlapped spectra. Each dataset contains its own series of noise. One dataset was used for training, 10 for validation and one as a test dataset. The training and validation datasets were used for the determination of the parameter λ, and the test set was used for final evaluation of algorithms performance. For the spectra, two Breit-Wigner peaks were generated with a relative separation of 0.35, defined in terms of the ratio between the distance between maxima and the FWHM (a value of 0 would indicate the peaks are completely overlapped; a value of 1 would indicate the peaks are resolved at half height). Regarding the ratio between the peak FWHM and the FWHM for a Gaussian of the model at the same position, for the current synthetic dataset this ratio is 1.51 (a value of 1 would indicate that the peak is perfectly Gaussian). The concentration profiles related to each peak were very similar in order to test the algorithms under challenging conditions. The angular distance between profiles, which provides an estimation of the difficulty of the problem, can be calculated as:

(Eq. 4.8)
$$\theta_{ij} = \arccos\left[\frac{C_i \cdot C_j}{\|C_i\| \|C_j\|} \right]$$

Where C_x is the vector containing the evolution in time for the selected spectrum (concentration profile) and $\|C_x\|$ is the Euclidean norm. An angular distance of 0º would indicate that the concentration profiles are identical. The angular distance was

24º for the training dataset, 18º for validation and 10º for test. The relative maximum intensities for the contributions were 1000 and 700 for training, 1000 and 800 for validation, and 1000 and 500 for test; therefore second peaks appear in the tail of the first ones.

In the case of LASSO, the tuning parameter λ is determined by cross-validation using the training and validation datasets. In order to prevent overfitting in the selection of λ, the training, validation and test datasets were created with different concentration profiles as described previously. Given a certain value for λ, the LASSO method is applied to the training dataset; the recovered spectra S_{train} are used to estimate the concentration profiles C_{val} using a validation dataset and FNNLS in a single step. The RMSE (Eq. 4.6) is evaluated between the validation dataset and the reconstructed dataset using C_{val} and S_{train}. An averaged RMSE is obtained for each λ value. High values of λ constrict too much the model; low values of λ allow noise appearance. Therefore, the best value for lambda is that around a change in the slope in the graph RMSE vs. λ (Figure 4.3). Usually, values within a range around the optimum provide similar results; thereby, the selection of the value is not critical. After selecting the λ value, this is applied to the test dataset to compare the MCR solutions.
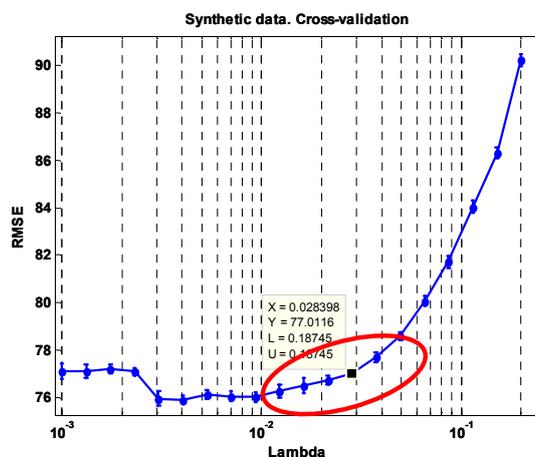
Figure 4.3. Determination of the tuning parameter λ by cross-validation. The λ value is selected as that around the change in the RMSE slope. Other possible values around the selected value are shown within the ellipse.

4.3.4 Ion Mobility Experiments with a Baggage Scanner

Experiments were performed using a baggage scanner provided by RAMEM S.A. (Madrid, Spain) (RAMEM). The scanner comprises a conveyor belt with a moving velocity of 0.5m/s and a chamber, located in the middle of the conveyor belt, with an approximate effective volume of 200l. Inside the chamber there are 8 compressed air nozzles. The compressed air has a pressure up to 6 bars and a flow up to 300l/min.

Furthermore, within the chamber, there is a 1000W infrared lamp. Both, the compressed air and the infrared lamp, can be switched on in order to favor the vaporization of volatiles present in the sample. The chamber's air is extracted using an extractor fan with a flow up to 77m^3/h which is located into a chimney connected with the chamber. In the chimney there is a connection with the instrument.

The measurements were performed using a gas detector array (GDA) manufactured by *Airsense Analytics GmbH* (Schwerin, Germany). The GDA includes an IMS with a radioactive ionization source of ^{63}Ni, it is based on water chemistry and its resolving power (Rokushika et al. 1985) is 32 at least (Eq. 1.3). The GDA incorporates an internal pump with a flow of 400ml/min. The IMS provides a different sample spectrum of length 28ms every 3 seconds. This spectrum corresponds to an average of 16 consecutive spectra for noise reduction. The sampling time for the drift spectra is 0.03ms. A picture of the instrument is shown in Figure 1.15(a).

The luggage scanner can be controlled using a specific software designed using LabView 8.5 (National Instruments, Texas, USA), provided by RAMEM. This software allows controlling the duty cycle of the compressed air nozzles, the time that a suitcase remains within the chamber and the time that the infrared lamp is switched on.

The sample was pipetted on a glass fiber substrate located over a suitcase with dimensions 30x15x20cm^3. The suitcase is introduced within the chamber using the conveyor belt. A picture of the mechanical system is shown in Figure 4.4.
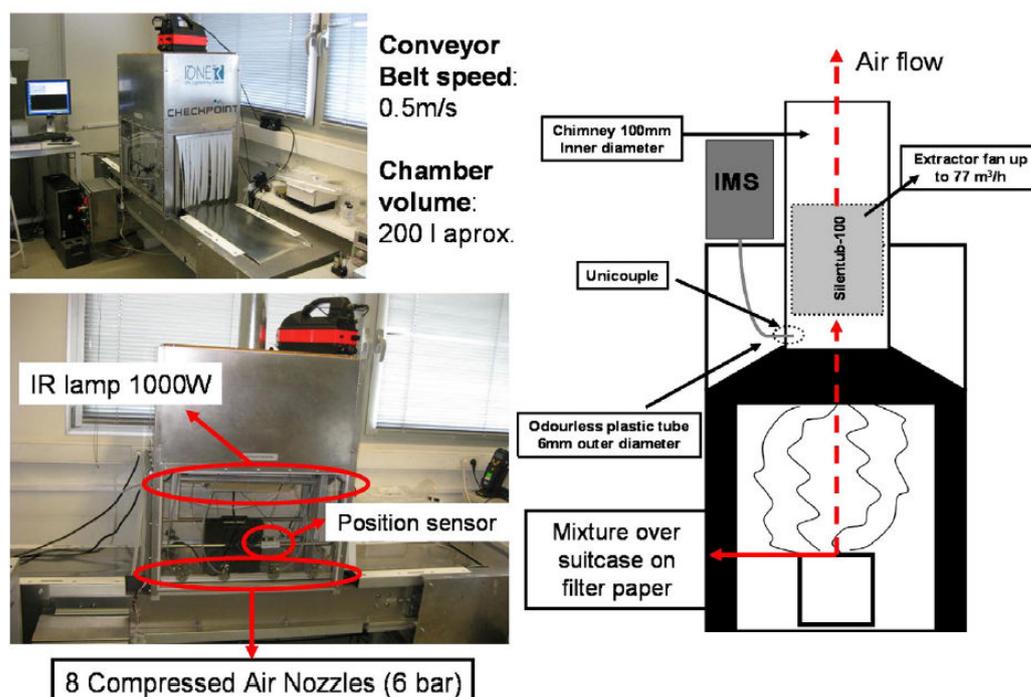
Figure 4.4. Scanner for baggage inspection. The schematic shows the locations of the IMS and the extractor fan.

Two experiments using the baggage scanner are presented in this chapter. In the first experiment, 10µl of benzaldehyde were pipetted on the glass fiber substrate. The sample was 6 seconds inside the chamber with the infrared lamp permanently turned on. In the second experiment, the sample consists of 0.5µl of ethanol and 0.5µl of ortho-nitrotoluene (o-MNT), a taggant for explosive detection (Ewing et al. 2001). The sample was 12 seconds inside the chamber with the infrared lamp and the air compressed nozzles permanently turned on.

The reader should notice that although the samples were in the µl range on the glass, the chamber had a volume of 200l and the extractor fan had a flow of 77m^3/h, therefore the concentration of chemicals arriving at the detector had been strongly diluted.

Three replicates (training, validation and test) per experiment were performed.

In the MCR-LASSO case, using the training and validation datasets, the regularization parameter λ is estimated as explained in section 4.3.3 and applied to the test dataset.

From our results, the tuning parameter λ depends on the width of the peaks and the noise level. Regarding the width of the peaks, the wider the peak is, the more number of Gaussians are needed to reconstruct the spectrum; thus λ needs to be small enough to not constrict too much the growth of the coefficients. Concerning the noise, the higher the noise level is, the higher the λ is needed for filtering purposes. Since the experiments were performed using the same instrument in similar operating conditions, it seems a reasonable assumption to fix a λ value for the IMS experiments. The parameter λ was estimated from the experiment with benzaldehyde. An appropriate value for λ has been found at 0.1. This value is used within MCR-LASSO in the two experiments (test datasets).

4.3.4.1   Baseline subtraction

Baseline from each spectrum was identified from the initial and final part of the spectrum where no peaks were identified. The first 150 points (from 1ms to 5.51ms) and the last 295 points (from 19.15ms to 28.09ms) were fitted to a polynomial of 4th order. Figure 4.5 shows and example of baseline correction for a spectrum from the dataset of the second experiment (similar results were obtained for the first experiment). The baseline correction was applied independently spectrum by spectrum.
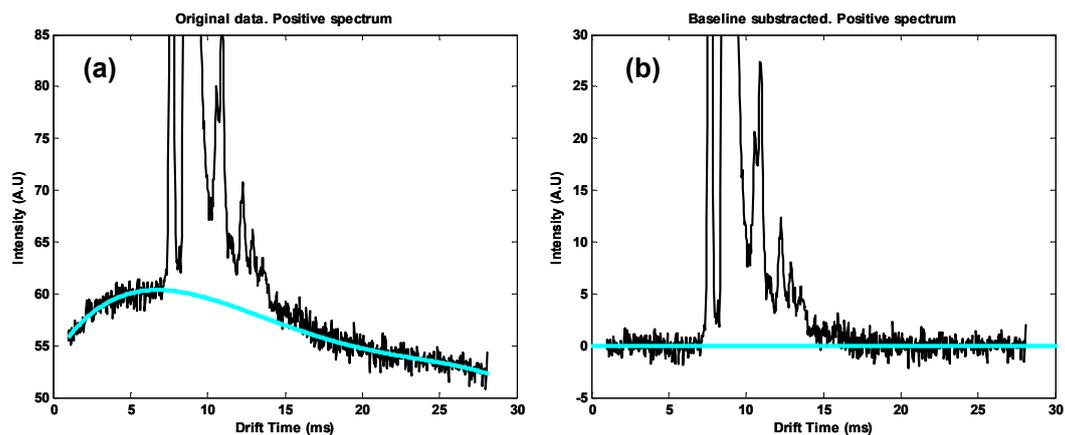
Figure 4.5. Example of baseline correction for the spectra recorded using the IMS. (a) Baseline for a spectrum from the dataset of the second experiment using the baggage scanner. (b) Spectrum after applying the 4th order polynomial correction.

4.3.5 Algorithms Implementation Details

MCR-LASSO was programmed using MATLAB 7.5 (*Mathworks*) and PLS toolbox 5.2 (*EigenVector Research*, Seattle, USA). The MATLAB software for MCR-ALS is available in (Tauler & Juan 2005). The algorithm used for SIMPLISMA corresponds to the MATLAB function *purity*, which can be found in PLS Toolbox 5.2. SIMPLISMA was applied using the pure variable approach in all the cases; the description of the function can be found in (Windig et al. 2005). For comparison, the MCR-LASSO will be compared with the MCR-ALS in exactly the same conditions regarding initialization and convergence criteria.

In the experiments presented in this chapter, only a fixed number of components and non-negativity were applied as constraints to MCR techniques since we consider that, in a real scenario, other constraints are non-sense. Non-negativity was applied through FNNLS in both techniques (MCR-ALS and MCR-LASSO). Spectra were normalized to unit area in each iterative step within the algorithms. The stop criterion (the same for both algorithms) was based on the RMSE (Eq. 4.6) between the data matrix filtered with PCA and the reconstructed matrix using *C* and *S*. At each iteration, the RMSE is calculated and the algorithms stop if the relative difference between iterations is less than 0.1%.

While some approaches to find the optimal number of pure components (*K*) have been reported (Windig & Guilment 1991; Windig et al. 2005; Buxton & Harrington 2001; Gourvénec, Massart & Rutledge 2002), in our case the choice has been based on the Singular Value Decomposition (SVD) of the data matrix and visual inspection of the estimations for *S* and *C* using SIMPLISMA.

4.3.6 Figures of merit

4.3.6.1 Synthetic dataset

For the synthetic dataset, since the underlying solutions for spectra and concentration profiles are known, the RMSE (Eq. 4.6) and an orthogonal distance (Eq. 4.8) can be calculated. Equation (Eq. 4.6) is used taking the matrix of real spectra (profiles) and the matrix of recovered spectra (profiles) instead of the data matrices (D and $\hat{D}$). (Eq. 4.8) is used assessing the orthogonal distance between the real spectrum (profile) and the recovered spectrum (profile) by the MCR algorithms. An orthogonal distance of 0º would indicate that the recovered spectrum (profile) is identical to the synthetic one. Moreover, the recovered power, which is related to the RMSE, can also be calculated (see next section).

4.3.6.2 Experimental datasets

For the experimental datasets, since the underlying solutions are not known, as figure of merit, we consider the power recovered by the algorithms regarding the original data matrix. The power of the original data matrix D can be calculated as $P_0 = \sum_{m=1}^{M} \sum_{n=1}^{N} D_{m,n}^2$ (squared Frobenius norm). From (Eq. 4.5), the power of the matrix of residuals can be derived: $P_{res} = \sum_{m=1}^{M} \sum_{n=1}^{N} E_{m,n}^2 = \sum_{m=1}^{M} \sum_{n=1}^{N} \left[D - C \cdot S^T \right]_{m,n}^2$. Then, the recovered power can be obtained as:

(Eq. 4.9) $$P_{rec}(\%) = \left(P_0 - P_{res} \right) / P_0 \cdot 100$$

4.4 Results

4.4.1 Results for synthetic dataset

Figure 4.6 shows a comparison of the results provided by MCR-ALS (Figure 4.6(a) and (c)) and MCR-LASSO (Figure 4.6(b) and (d)) after selecting the first estimations from SIMPLISMA. The regularization parameter λ for MCR-LASSO was selected as explained in section 4.3.3 (Figure 4.3).
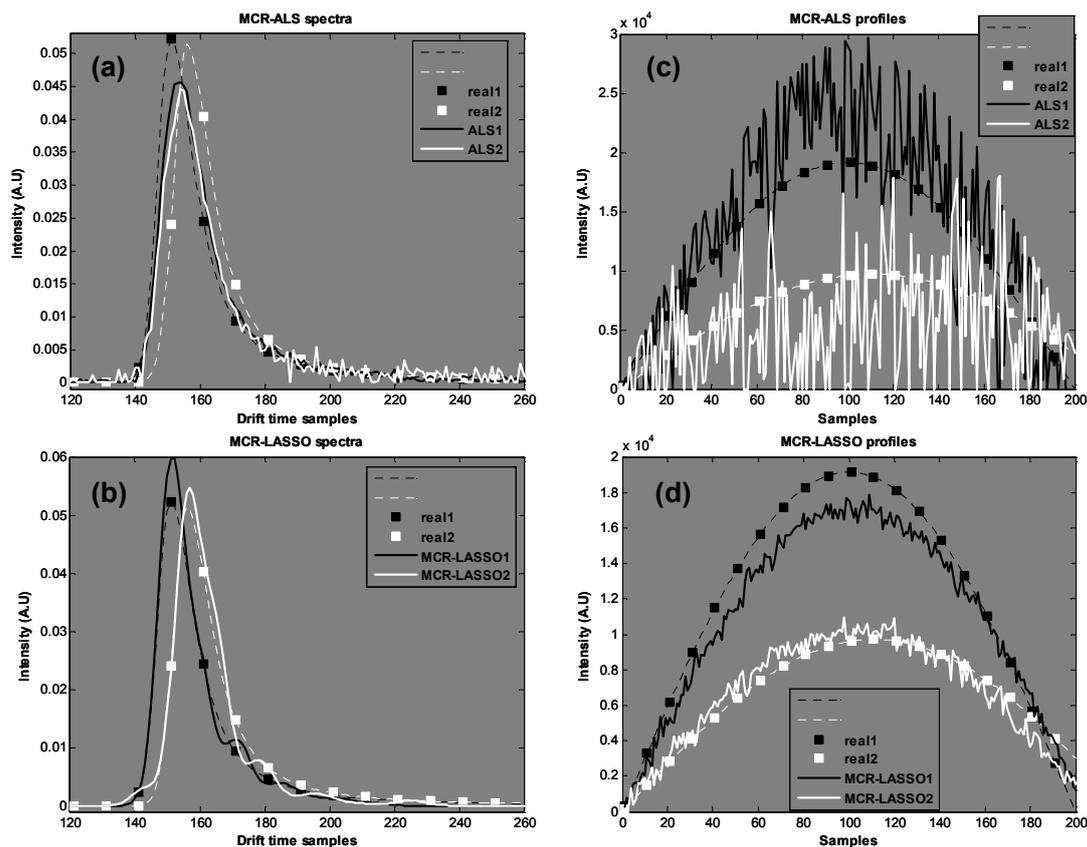
Figure 4.6. Results for synthetic dataset (SNR=10dB) with asymmetrical peaks (half width at half height left part w_g=4.7 and half width at half height right part w_{bw}=9.5) compared to MCR solutions. (a) Spectra recovered by MCR-ALS. (b) Spectra recovered by MCR-LASSO (λ=0.03). (c) Concentration profiles by MCR-ALS. (d) Concentration profiles by MCR-LASSO (λ=0.03).

Due to the challenging conditions, the first estimations from SIMPLISMA were very noisy in the concentration profiles and showing bad resolution in the spectra (not shown). MCR-ALS is not able to recover the peaks in the correct position (Figure 4.6 (a)) and the shape of the peaks is distorted due to noise. Moreover, the estimations for the concentration profiles are very noisy (Figure 4.6(c)). On the contrary, for a certain range of λ values, although the shapes of the peaks are a little bit distorted due to the rigidity of the Gaussian model, MCR-LASSO is able to recover properly the peak position for the spectra (Figure 4.6(b)) and provides stable, regularized and much less noisy estimations for the profiles (Figure 4.6(d)). In that sense, for suitable values of λ, MCR-LASSO is performing as a filter at the same time that regularizes the solution.

	RMSE (A.U)		Angle (º)		Power (%)
	Spectra	Profiles	Spectra	Profiles	
MCR-ALS	$2.6 \cdot 10^{-3}$	$4.9 \cdot 10^{3}$	10, 19	14, 40	91
MCR-LASSO	$1.2 \cdot 10^{-3}$	$1.2 \cdot 10^{3}$	4, 6	3, 6	91

Table 4.1. Comparison of quantitative results obtained with the synthetic dataset. Figures of merit are RMSE (Eq. 4.6), orthogonal angle (Eq. 4.8) and recovered power (Eq. 4.9).

The quantitative results are presented in Table 4.1. It is shown that MCR-LASSO provides better solutions in terms of orthogonality (Eq. 4.8) and RMSE (Eq. 4.6). Regarding the recovered power (Eq. 4.9), both MCR-ALS and MCR-LASSO recover 91% of the total variance for the original noisy dataset.

4.4.2 Results for experimental datasets from the luggage scanner

4.4.2.1 Experiment 1: Benzaldehyde

This experiment shows a clear example of non-linearity; for high concentrations the main peak shifts to a longer drift time, maybe due to dimering. This leads to a dramatic spectrum distortion depending on concentration. This experiment will show that non-linearity can be dealt properly using additional components in the model. Ultimate correct model interpretation will depend on the criteria and previous knowledge of the human observer.
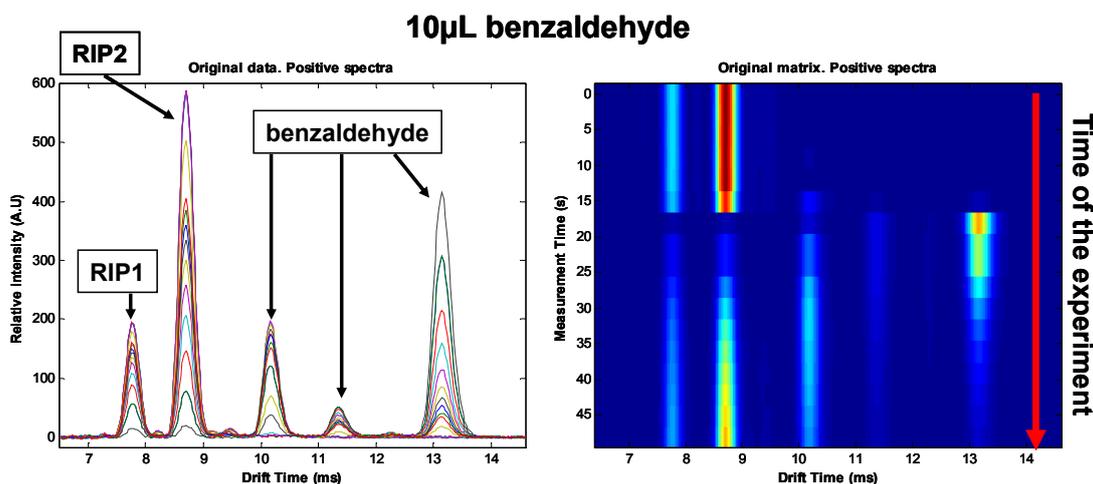
Figure 4.7. Spectra collection obtained using the IMS for the experiment with benzaldehyde after baseline correction. Left: Intensity vs. drift time. Right: Time of the experiment vs. drift time.

The data matrix obtained from the IMS after baseline correction is shown in Figure 4.7 (test dataset). The matrix comprises 17 spectra with 271 drift time points from 6.42ms

to 14.60ms, where relevant peaks appear. The total time of the experiment was 48 seconds. Figure 4.7 shows the peaks related to each substance and their drift time. The first two peaks are the reactant ion peaks. Figure 4.7 also shows the three peaks related to benzaldehyde, which have a different evolution in time.

First estimations for spectra and profiles are obtained using SIMPLISMA (Figure 4.8(a) and (d)) selecting 5 pure components. Negative values appear in SIMPLISMA, which do not have physical meaning. Moreover, the fourth and fifth components in spectra are very noisy and contain contributions in regions where other peaks appear, trying to compensate its evolution.
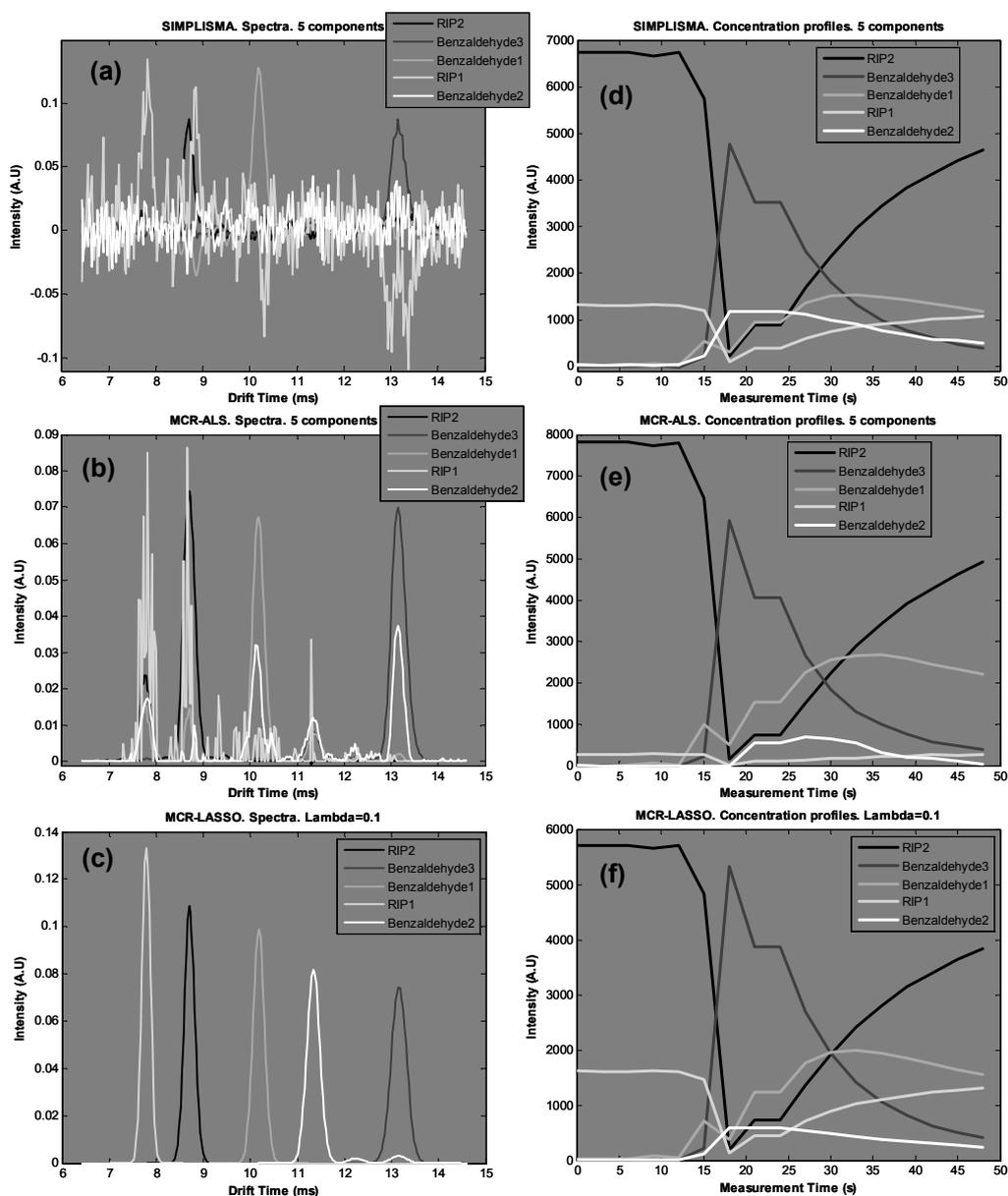
Figure 4.8. Results for SIMPLISMA and MCR approaches for the experiment with benzaldehyde imposing 5 components. Left (spectra): (a) SIMPLISMA, (b) MCR-ALS, (c) MCR-LASSO (λ=0.1). Right (concentration profiles): (d) SIMPLISMA, (e) MCR-ALS, (f) MCR-LASSO (λ=0.1).

Figure 4.8(b) shows the spectra recovered by MCR-ALS imposing non-negativity, therefore no negative values appear. The component spectra have contributions in regions where other peaks appear for other spectra (peak overlapping). This is clearly visible in the fourth and fifth components, which try to model the behaviour for the first RIP and the second benzaldehyde peak, respectively. Moreover, these components are very noisy due to the estimation introduced by SIMPLISMA for them, thus MCR-ALS is not able to provide an easy interpretation for the behaviour of these variables. This fact hinders the interpretation of the results.

Figure 4.8(c) shows the spectra recovered by MCR-LASSO with $\lambda=0.1$ (selected from cross-validation). The noise in the spectra has been highly reduced and practically there are no contributions of one spectrum in regions where peaks appear for other spectra. Even though the noisy estimation for the fourth and fifth components, the algorithm is able to regularize and stabilize the solution. This fact makes easier the interpretation of the concentration profiles related to each spectrum (Figure 4.8(f)). When the sample is introduced into the chamber transference of charge, from the reactant ion peaks to the incoming molecules, occurs. Three peaks are related to benzaldehyde and they have a different evolution in time depending on the concentration of the substance. The first two peaks appear at low concentrations, while the third peak appears only at higher concentrations. After sample introduction, the first two peaks appear but very fast charge is transferred to the third peak. When concentration decreases, the third peak disappears and the two initial peaks gain importance only to finally disappear again when benzaldehyde solution goes to zero. When the sample is extracted from the chamber, the reactant ion peaks begin to recover its charge at the same time that the concentration of benzaldehyde decreases in intensity. This variation in concentration is complex from the chemical point of view of the IMS and it is reflected in the different evolution of the benzaldehyde peaks.

In terms of recovered variance (Eq. 4.9) by the different algorithms with regard to the original data matrix, SIMPLISMA and MCR-ALS provide almost 100% of the original variance; MCR-LASSO 99%. Despite that MCR-LASSO is providing less variance than the other approaches, the recovered power is still highly significant. This decreasing in variance is attributed to the rigidity of the Gaussian model.

4.4.2.2 Experiment 2: Ethanol + o-MNT + interferences

This second experiment provides a good example to see how the algorithms perform in the presence of interference substances, thereby in a more complex mixture. The data matrix obtained from the IMS after baseline correction is shown in Figure 4.9 (test dataset). The matrix comprises 91 spectra with 301 drift time points from 5.51ms to 14.60ms, where relevant peaks appear. The total time of the experiment was 273

seconds. Figure 4.9 shows the peaks related to each substance present in the mixture and their drift time. The first two peaks are the reactant ion peaks (RIP). Figure 4.9 also shows the peak related to ethanol and the three peaks related to o-MNT. Moreover, a peak due to interfering chemicals appears overlapped with the peak of ethanol (according to our interpretation, this interference is related to compressed air since it disappears if air compressed flow is turned off) and also there is an acetone remainder from previous experiments.
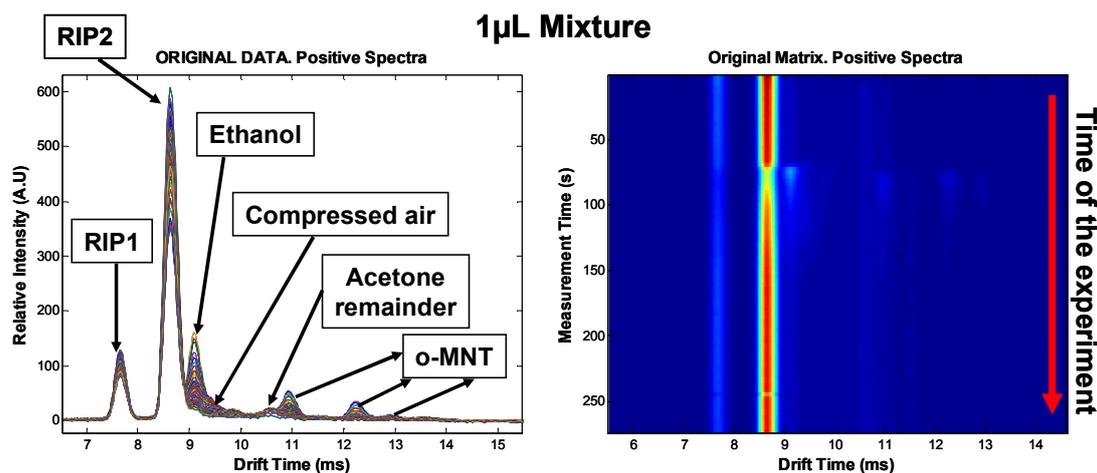
Figure 4.9. Spectra collection obtained using the IMS after baseline correction for the mixture of ethanol and o-MNT in the presence of interferences. Left: Intensity vs. drift time. Right: Time of the experiment vs. drift time.

First estimations for spectra and profiles are obtained using SIMPLISMA. Figure 4.10 shows the results provided by SIMPLISMA for five (Figure 4.10(a), (b) and (c)) and six components (Figure 4.10(d), (e) and (f)). For five components, a peak with very high intensity appears in the resolved spectrum due to interfering chemicals present in compressed air (Figure 4.10(a)). This peak with high intensity tries to compensate negative values for this contribution in other drift time intervals. Moreover, negative values appear in the spectra (Figure 4.10(a) and (d)), which do not have physical meaning. Concerning the concentration profile, for the interfering contribution (Figure 4.10(c)), this is close to zero. Thus, results provided by SIMPLISMA seem to be bad in these cases but still they provide a good initial point for the bilinear matrix decomposition (Eq. 4.5).
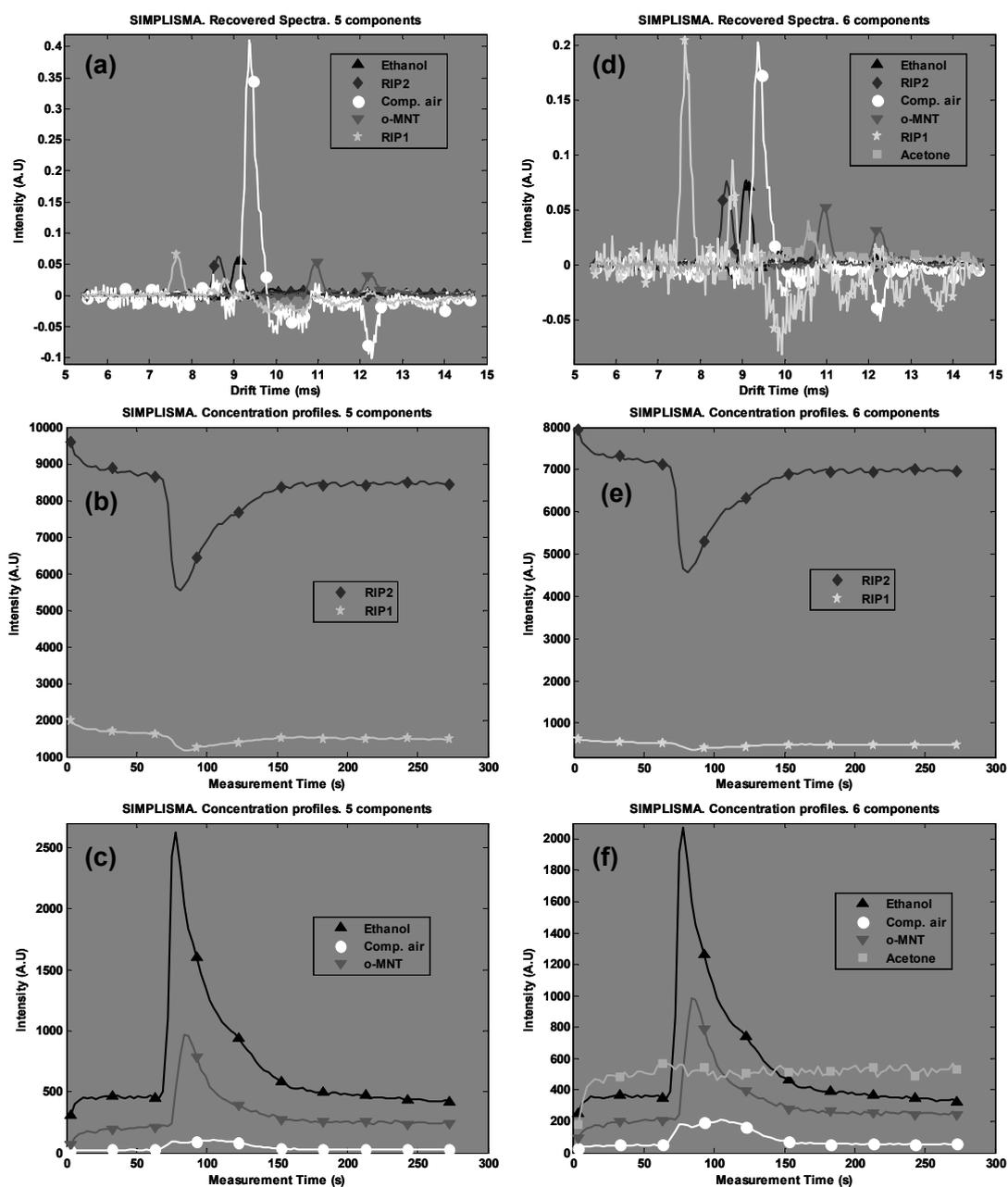
Figure 4.10. First estimations for MCR approaches using SIMPLISMA. Left (5 components): (a) Spectra, (b) Profiles for the two reactant ion peaks, (c) Profiles for ethanol, compressed air and o-MNT. Right (6 components): (d) Spectra, (e) Profiles for the two reactant ion peaks, (f) Profiles for ethanol, compressed air, o-MNT and acetone remainder.

Figure 4.11 and Figure 4.12 show the spectra and concentration profiles recovered by MCR-ALS and MCR-LASSO (λ=0.1) imposing five and six components, respectively.
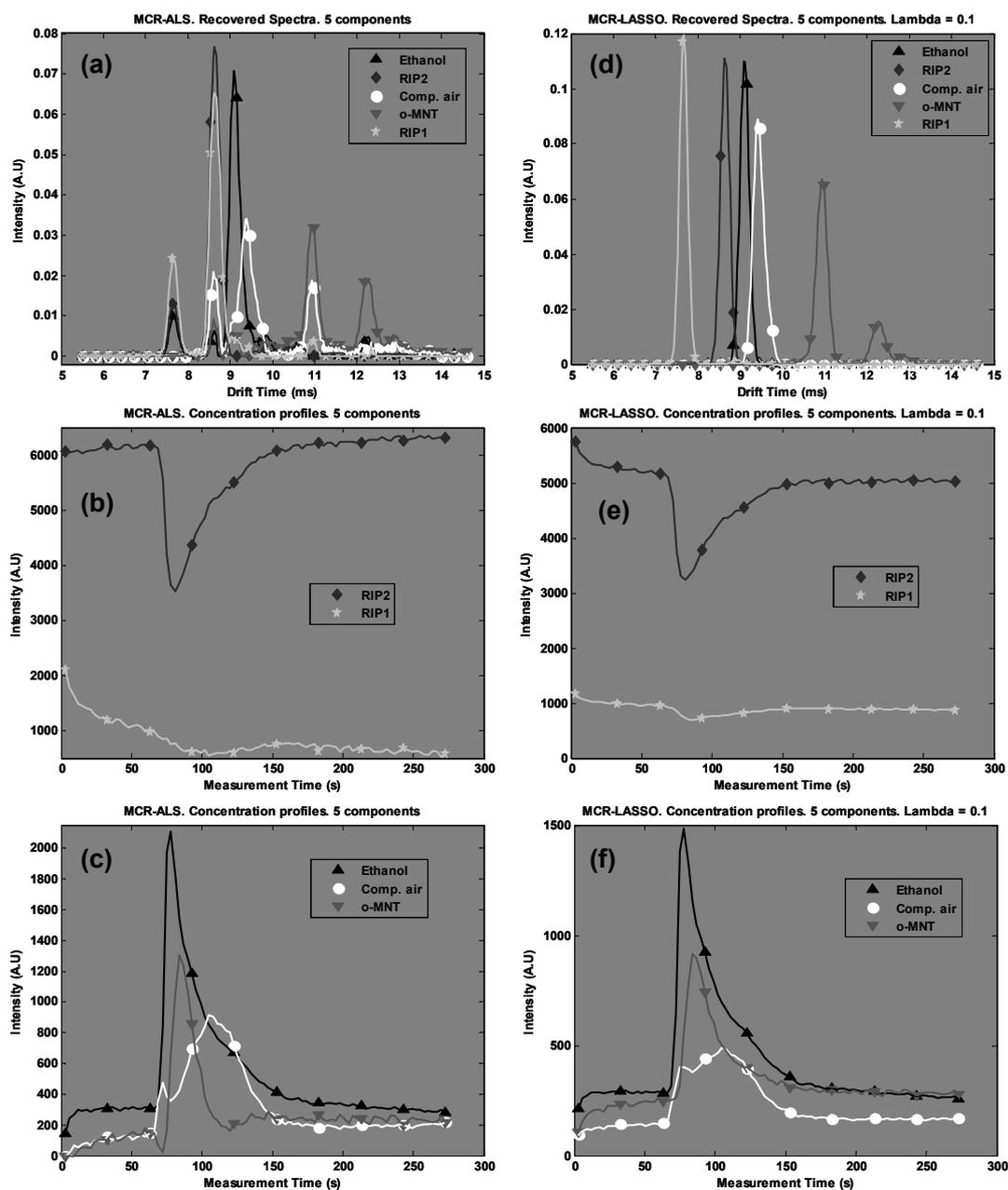
Figure 4.11. Spectra and concentration profiles recovered imposing 5 pure components using MCR-ALS (left) and MCR-LASSO with *λ*=0.1 (right). Non-negativity constraint was applied through FNNLS. (a) MCR-ALS spectra, (b) MCR-ALS reactant ion peaks profiles, (c) MCR-ALS profiles for ethanol, compressed air and o-MNT, (d) MCR-LASSO spectra, (e) MCR-LASSO reactant ion peaks profiles, (f) MCR-LASSO profiles for ethanol, compressed air and o-MNT.

Concerning the case with five components (Figure 4.11), Figure 4.11(a), (b) and (c) show the results provided by MCR-ALS. A high degree of overlapping can be observed among the recovered spectra related to each pure component (Figure 4.11(a)). Regarding concentration profiles (Figure 4.11(b)), at the beginning of the experiment (measurement time), the pure components associated to the two RIPs are trying to

compensate to each other. This effect can not be explained taking into account the transference of charge from the RIPs to the incoming molecules in the IMS. Then, MCR-ALS is not separating properly the fraction of charge transfer due to each RIP.

Figure 4.11(d), (e) and (f) show the results provided by MCR-LASSO. Compared to MCR-ALS, overlapping is smaller among spectra, then, MCR-LASSO results can be interpreted in an easier manner (Figure 4.11(d)). Moreover, signal-to-noise ratio (SNR) is significantly improved in MCR-LASSO due to the hard constraint imposed by the IMS Gaussian model. Concerning concentration profiles (Figure 4.11(c) and (f)), both algorithms are providing similar results.

As can be seen in Figure 4.11(e), when the sample is introduced into the chamber, the two RIPs decrease in intensity due to the transference of charge to the incoming molecules, and the intensity of the peaks related to ethanol, compressed air and o-MNT increase (Figure 4.11(f)). Ethanol, with a higher volatility, appears first in time at 69s and it reaches its maximum intensity at 78s because it has fully evaporated from the fiber glass substrate. o-MNT appears secondly at 72s and it begins to decrease at 84s. This is in perfect agreement with the time that the suitcase remained inside the chamber (12s). Compressed air nozzles were turned on a few seconds before the suitcase entered within the chamber, and they were turned off 30s later then, concerning the interference due to compressed air, it appears at 66s (before the appearance of ethanol) and grows slowly in intensity until it reaches a maximum value at 105s (39s), which is in agreement with the experimental conditions. Exactly the same analysis can be done analyzing Figure 4.11(c), although the shapes of the curves are slightly different from Figure 4.11(f).

An extra component can be introduced in the algorithms in order to model the behaviour of the acetone remainder (Figure 4.12). The analysis of the results is very similar to the previous case (Figure 4.11) with slight differences. Regarding concentration profiles (Figure 4.12(c) and (f)), MCR-ALS is not modelling correctly the evolution of acetone, which increases significantly its intensity in time. MCR-LASSO provides a better result but still gives a slight increase in intensity for the component. Only a decreasing in intensity should be expected since the component was not present in the mixture and is a remainder from previous experiments. MCR-LASSO is still providing better results in terms of overlapping among spectra and signal-to-noise ratio (Figure 4.12(d)), which makes easier the interpretation of the concentration profiles.

Although MCR-LASSO captures a little less variance than MCR-ALS (almost 100% for five and six components), basically due to the rigidity of the Gaussian model plus the sparse solution, the recovered power (Eq. 4.9) is still highly significant (99% for five and six components).
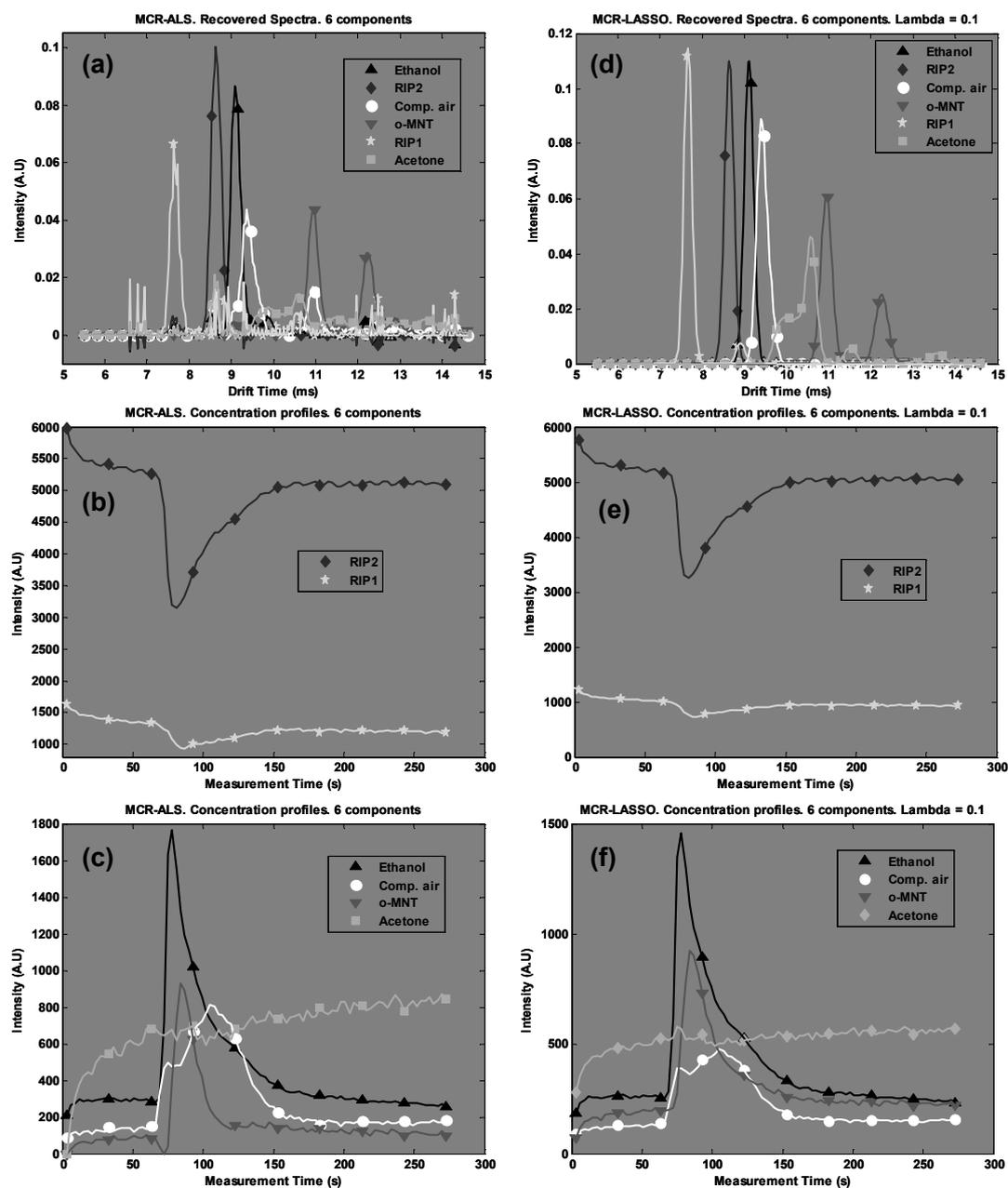
Figure 4.12. Spectra and concentration profiles recovered imposing 6 pure substances using MCR-ALS (left) and MCR-LASSO with λ=0.1 (right). Non-negativity constraint was applied through FNNLS. (a) MCR-ALS spectra, (b) MCR-ALS reactant ion peaks profiles, (c) MCR-ALS profiles for ethanol, compressed air, o-MNT and acetone remainder, (d) MCR-LASSO spectra, (e) MCR-LASSO reactant ion peaks profiles, (f) MCR-LASSO profiles for ethanol, compressed air, o-MNT and acetone remainder.

4.5 Conclusions

It is well-known that imposing constraints in MCR-ALS leads to better solutions and results that are easier to interpret qualitatively. However, when using ion mobility spectrometry, some constraints are not applicable in a real scenario where interfering chemicals are expected to be found, since the spectra could be very complicated. In these cases, integrating as much knowledge about the system as possible could be critical to detect certain chemicals.

We have presented MCR-LASSO, a technique for multivariate curve resolution that introduces a flexible model for the spectra in the form of a dense superposition of Gaussian peaks. The LASSO technique is introduced within the ALS loop for spectra modelling. Model complexity is limited by the use of L1-norm regularization.

Synthetic experiments have shown that in challenging conditions (high noise, very similar concentration profiles, overlapped spectra, and asymmetric peaks) MCR-LASSO provides better estimation of the time evolution and spectra of the underlying components. The dense superposition of Gaussian is able to model wider asymmetric peaks usually found in spectroscopy.

On the other hand, MCR-LASSO has been found also to provide better resolution in two real experiments using a baggage scanner prototype. The first has shown that IMS non-linearities can be dealt by the introduction of additional pure components. The second experiment presents a more complex mixture including the presence of interfering chemicals. MCR-LASSO results contain less noise, not only in the spectra but also in the concentration evolution, and their interpretation is easier than that from MCR-ALS or SIMPLISMA. At this point, it is also important to mention that the LASSO approach can be used with other peak models apart from the Gaussian one chosen in this work.

The use of a regularized solution using the L1-norm permits to use a flexible model (in this case a very dense superposition of Gaussian peaks) that would otherwise result in a very ill-conditioned least squares problem, and still find a sparse solution of limited complexity.

Although our work has been based on IMS spectra and Gaussian peaks, we believe that MCR-LASSO can be applied to other analytical settings provided that spectra models can be based on dense linear superposition of regressors.

4.6 References

Bro, R & Jong, Sd 1997, 'A fast non-negativity-constrained least squares algorithm', *Journal of Chemometrics*, vol. 11, pp. 393-401.

Bro, R & Sidiropoulos, ND 1998, 'Least Squares algorithms under unimodality and non-negativity constraints', *Journal of Chemometrics*, vol. 12, pp. 223-247.

Buxton, TL & Harrington, PdB 2001, 'Rapid multivariate curve resolution applied to identification of explosives by ion mobility spectrometry', *Analytica Chimica Acta*, vol. 434, pp. 269-282.

Chen, SS, Donoho, D & Saunders, M 1998, 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33-61.

Donoho, D 2006, 'Compressed sensing', *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306.

Eldén, L 1977, 'Algorithms for the regularization of ill-conditioned least squares problems', *BIT Numerical Mathematics*, vol. 17, no. 2, pp. 134-145.

Ewing, RG, Atkinson, DA, Eiceman, GA & Ewing, GJ 2001, 'A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds', *Talanta*, vol. 54, pp. 515-529.

Fan, J & Li, R 2001, 'Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties', *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348-1360.

Felinger, A 1998, 'Peak Shape analysis', in *Data analysis and signal processing in chromatography*, vol. 21, Elsevier Science, pp. 97-124.

Gemperline, PJ & Cash, E 2003, 'Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions', *Analytical Chemistry*, vol. 75, no. 16, pp. 4236-4243.

Golub, G 1965, 'Numerical methods for solving linear least squares problems', *Numerische Mathematik*, vol. 7, no. 3, pp. 206-216.

Gourvénec, S, Massart, DL & Rutledge, DN 2002, 'Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the Orthogonal Projection Approach and in the SIMPLe-to-use Interactive Self-modelling Mixture Analysis approach', *Chemometrics and Intelligent Laboratory Systems*, vol. 61, no. 51-61.

Hastie, T, Tibshirani, R & Friedman, J 2009, 'Linear Methods for Regression', in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edn, Springer Series in Statistics, pp. 43-100.

Jaumot, J & Tauler, R 2010, 'MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution', *Chemometrics and Intelligent Laboratory Systems*, vol. 103, no. 2, pp. 96-107.

Juan, Ad, Maeder, M, Martinez, M & Tauler, R 2000, 'Combining hard and soft modelling to solve kinetic problems', *Chemometrics and Intelligent Laboratory Systems*, vol. 54, pp. 123-141.

Juan, AD, VanderHeyden, Y, Tauler, R & Massart, DL 1997, 'Assessment of new constraints applied to the alternating least squares method', *Analytica Chimica Acta*, vol. 346, pp. 307-318.

RAMEM, *http://www.ioner.eu/*, RAMEM S.A.

Rokushika, S, Hatano, H, Bairn, MA & Jr., HHH 1985, 'Resolution Measurement for Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 57, no. 9, pp. 1902-1908.

Siems, WF, Wu, C, Tarver, EE, Hill, HH, Larsen, PR & McMinn, DG 1994, 'Measuring the resolving power of ion mobility spectrometers', *Analytical Chemistry*, vol. 66, no. 23, pp. 4195-4201.

Spangler, GE 2002, 'Expanded theory for the resolving power of a linear ion mobility spectrometer', *International Journal of Mass Spectrometry*, vol. 220, pp. 399-418.

Tauler, R 1995, 'Multivariate curve resolution applied to second order data', *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 133-146.

Tauler, R & Juan, Ad 2005, *http://www.ub.edu/mcr/ndownload.htm* in Multivariate Curve Resolution Homepage.

Tibshirani, R 1996, 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267-288.

Vogtland, D & Baumbach, JI 2009, 'Breit-Wigner-Function and IMS-signals ', *International Journal for Ion Mobility Spectrometry*, vol. 12, no. 3, pp. 109-114.

Wang, J-H, Hopke, PK, Hancewicz, TM & Zhang, SL 2003, 'Application of modified alternating least squares regression to spectroscopic image analysis', *Analytica Chimica Acta*, vol. 476, pp. 93–109.

Windig, W, Gallagher, NB, Shaver, JM & Wise, BM 2005, 'A new approach for interactive self-modeling mixture analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, pp. 85-96.

Windig, W & Guilment, J 1991, 'Interactive Self-Modeling Mixture Analysis', *Analytical Chemistry*, vol. 63, no. 14, pp. 1425-1432.

Chapter 5

Qualitative Analysis and Quantitative Prediction of Non-Linear Ion Mobility Spectra

5.1 Introduction

In this chapter, a new methodology to analyze spectra time-series obtained from ion mobility spectrometry (IMS) is investigated with the ultimate purpose of quantitative prediction. Quantitative prediction in IMS is hindered by the presence of strong non-linearities due to the appearance of dimers or even more complex clusters of ions. Those non-linearities are particularly important in the presence of humidity since ions can be protonated at different levels. In the presence of mixtures, mixed ions cluster are also possible (Puton et al. 2012). In the current chapter, the analyzed IMS spectra present a strong non-linear behaviour as substance concentration increases (section 1.3.1 and Figure 5.1).

The proposed method combines the advantages of multivariate curve resolution-alternating least squares (MCR-ALS) for an optimal physical and chemical interpretation of the bilinear decomposition (Eq. 1.9) of the data matrix (qualitative information) and a multivariate calibration technique such as polynomial partial least squares (poly-PLS) (Wold, Kettaneh-Wold & Skagerberg 1989; Rosipal 2008) for a final quantification (quantitative information) of new samples. This method allows coping with cases where spectrum behaviour is non-linear with the concentration.
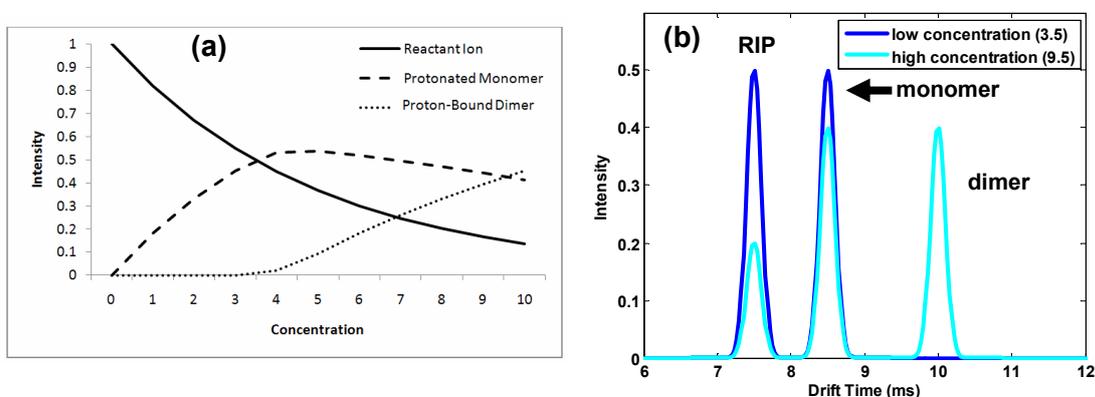
Figure 5.1. Synthetic representation of the effect of concentration response in traditional ion mobility spectrometry. (a) Concentration profiles as concentration increases. (b) Spectral responses for two particular concentrations.

In order to build a predictive calibration model in IMS, univariate techniques appear as the first solution. A calibration could be performed using the information of peak area or height of protonated monomer or protonated-bound dimer and then applying an appropriate fitting function. Typically, fitting functions are polynomial. However, such techniques are not useful if peaks are overlapped (unless a previous deconvolution is done). While it is possible to keep the instrument in a linear regime for low input concentrations, this seriously damages the dynamic range (ratio of maximum concentration to the limit of detection) of the instrument for most analytes. Moreover, if peaks behaviour is non-linear as concentration is increased, quantitation is degraded if only one of the peaks is used. As it can be seen in Figure 5.1, a calibration model based on protonated monomer is very sensitive at low concentrations and provides no information at high concentrations; on the other hand, a calibration model based on protonated-bound dimer is quite the opposite. Therefore, multivariate calibration techniques appear to be a good choice dealing with non-linearities and peak overlapping.

Only in the last twenty years multivariate calibration models have been proposed to process IMS for quantitative analysis (Zheng, Harrington & Davis 1996; Fraga, Kerr & Atkinson 2009; Ochoa & Harrington 2005; Zamora & Blanco 2012; Lu, O'Donnell & Harrington 2009). Recently, Fraga et al. in (Fraga, Kerr & Atkinson 2009) compared the performance of PLS and PCR with univariate regression based on peak area in the quantification of TNT and RDX in explosives. It is shown that multivariate calibration methods provide better IMS quantitative precision and accuracy than univariate methods even when the peaks are resolved.

Also recently, Zamora et al. (Zamora & Blanco 2012) demonstrate that MCR can be effectively applied to improve the resolution of overlapped peaks in IMS. In this work, two active principal ingredients (API) at low concentrations are also successfully quantified using PLS models.

While previous works have shown that PLS can be applied directly for quantitative prediction in IMS spectra, the interpretation of PLS models for IMS is not easy. PLS attempts to find factors which both capture variance of the spectral responses and achieve correlation between the spectral responses and the magnitude to be predicted (concentrations). However, both loadings and final regression coefficients can have negative values leading to compensation effects which increase even more the difficulty of physical or chemical interpretation. Furthermore, despite the fact that PLS algorithm is able to handle slightly non-linear data by increasing the number of latent variables in the calibration model, this approach is less successful for datasets containing moderate and severe non-linearities (Yang, Griffiths & Tate 2003). Several

variants of linear PLS have been developed to deal with non-linear datasets (Rosipal 2008), for instance using splines (Wold 1992) or using a polynomial of a certain order (Wold, Kettaneh-Wold & Skagerberg 1989; Baffi, Martin & Morris 1999). However, loadings remain difficult to interpret.

A different approach to analyze IMS spectra is the use of multivariate curve resolution techniques (MCR) that aim to recover the evolution of the source signals (concentration profiles) and the mixing matrix (spectral features) without any prior supervised calibration step. We use MCR with alternating least squares (MCR-ALS) (Tauler, Kowalski & Fleming 1993). The bilinear decomposition obtained leads to a feasible physical and chemical interpretation of the results. Since MCR-ALS is a linear method, nonlinear contributions can be modelled adding extra components.

As it is shown in Figure 5.1, since a single analyte could produce more than one ionic species with different evolution as concentration increases, we would need additional dimensions to model the variation of the overall spectra. In general it is reasonable to think that as many dimensions are needed as different ions appear.

Although MCR-ALS has been demonstrated in many different applications to resolve properly different contributions present in a sample, it only provides qualitative information. Since no prior knowledge about the composition of the mixture is assumed, MCR-ALS does not provide concentration information and cannot be used directly for quantification of new samples in its basic form.

However, posterior quantitation on the basis of MCR decomposition has been previously described. Antunes et al. (Antunes et al. 2002) and later Azzouz and Tauler presented the correlation constraint in (Azzouz & Tauler 2008). This constraint is introduced within the main ALS loop and provides a way to quantify new samples using MCR-ALS. However, the correlation constraint cannot be applied when different ion species produced by the same analyte are modelled as separated components, since only the analyte concentration is available and not the concentration of each ion species (Figure 5.1).

Specifically, for the analysis of traditional IMS spectra where the same analyte produces several ion peaks, Harrington et al. (Harrington & Chen 2004) proposes to use a set of equilibrium equations to model the relative concentration of the different ions in the drift tube. However, this approach is very dependent on the validity of the equilibrium model. If the model fit fails (e.g. because equilibrium conditions are not reached), the quantitative prediction accuracy degrades.

It is also shown in (Armenta & Blanco 2012) the effective use of MCR-ALS to obtain qualitative and quantitative individual component information from overlapped peaks. In this application, active pharmaceutical ingredients (APIs) are monitored in the air of

a workplace in a pharmaceutical industrial site using IMS. Qualitative profiles are obtained using MCR-ALS and they are averaged and interpolated to the corresponding calibration curve in order to obtain a quantitative measurement. However, this methodology can only be applied when only one peak is related to each component (slight non-linearities).

In the present work, we aim to provide a multivariate calibration method for IMS spectra combining the advantages of MCR-ALS for qualitative interpretation and a non-linear multivariate technique such as poly-PLS (Wold, Kettaneh-Wold & Skagerberg 1989; Rosipal 2008) for an improved quantification of substance concentration. Thereby, MCR-ALS is used as a prior step to multivariate calibration modelling nonlinear contributions properly and with an easier interpretation. This method can be useful especially in cases where peak intensity behaviour is non-linear as concentration increases.

5.2 Materials

5.2.1 Ion Mobility Spectrometer

In the present work, a handheld [63]Ni-based ion mobility spectrometer (GDA2, *Airsense Analytics*) is used (Figure 1.15(a)). The IMS is based on water chemistry, incorporates an internal pump with a flow of 400ml/min and provides a different sample spectrum of length 28ms every 3 seconds. This spectrum corresponds to an average of 16 consecutive spectra for noise reduction. The sampling frequency for the drift spectra is 33 KHz and the temperature of the drift tube reaches 45ºC in operating conditions. The IMS operated in positive ion mode.

For identification purposes the reduced mobility (K_0) is often used (Eq. 1.2), which allows comparing results between different IMS instruments using the same ionization source.

5.2.2 Sample preparation

Controlled and calibrated concentrations of 2-butanone and ethanol (at least 99% pure, provided by *Sigma-Aldrich*, St. Louis, USA) were prepared using synthetic air premier (pure at 99.995%, provided by *Carburos Metálicos*, Spain). The concentrations were obtained by dilution of the stream exiting the oven of a volatile generator system based on permeation tubes (OVG4, *Owlstone*, Cambridge, UK). The permeation tubes were previously calibrated in our facilities by gravimetric methods after one week in the OVG4 at constant temperature and flow.

The analytes were measured using the ion mobility spectrometer at ten different concentrations. Table 5.1 shows the measured concentrations for each substance. A certain number of spectra (stable consecutive scans) are obtained for each concentration. The same measurements were replicated in three different days. In the case of 2-butanone, the total size of the data matrix is 108 scans x 198 spectral points for each replicate. In the case of ethanol, the total size of the data matrix is 196 scans x 198 spectral points for each replicate.

2-butanone (ppm)	Ethanol (ppm)
0	0
0.05	0.39
0.10	0.89
0.16	1.5
0.23	2.1
0.29	2.7
0.33	3.0
0.38	3.6
0.44	4.1
0.51	4.7
0.57	5.3

Table 5.1. Concentrations for 2-butanone and ethanol generated using the volatile generator system.

5.3 Methods

5.3.1 Pre-processing of IMS datasets

Performing an appropriate spectral pre-processing in IMS data is critical to obtain reliable results. Three pre-processing steps have been applied to data matrices prior to MCR-ALS: baseline correction, noise reduction and spectra alignment. As it was shown in (Pomareda et al. 2010), baseline from each spectrum can be corrected fitting and subtracting a polynomial of 4th order using the initial and final part of each spectrum. In particular, the first 150 points (from 1ms to 5.51ms) and the last 295 points (from 19.15ms to 28.09ms) of each spectrum where no relevant peaks were identified (Figure 4.5). Additionally, noise reduction was performed using Savitzky-Golay filtering (Savitzky & Golay 1964) using a polynomial of second order of window length fifteen samples and finally, misalignment was corrected applying a shifting in drift time taking

as reference the position of the reactant ion peak (RIP). The last step is crucial in order to obtain good qualitative and quantitative results.

5.3.2 Qualitative analysis: MCR-ALS

MCR-ALS is used for a qualitative analysis of IMS datasets. The algorithm is described in chapter 1, 1.4.2.3. In each iterative step, non-negativity, unimodality and closure are used as constraints within the main ALS loop after computing (Eq. 1.10) or (Eq. 1.11). Non-negativity has been used because concentration profiles and spectra are expected to be positive in order to have a physical and chemical meaning. This constraint has been applied through fast nonnegative least squares (FNNLS)(Bro & Jong 1997). Moreover, unimodality has been applied in peaks which were expected to be unimodal (for instance, the reactant ion peak and the monomer). The closure constraint has also been applied because, in IMS with APCI (Atmospheric pressure chemical ionization), available charge is transferred among ions but this charge remains constant during the whole process; this constraint is applied to all concentration profiles.
Initial estimations for MCR-ALS are obtained using SIMPLISMA (Windig et al. 2005).

5.3.3 Quantitative analysis: univariate and multivariate calibration

Although MCR-ALS provides a powerful way to resolve different contributions measured in the spectra, the results are only qualitative. Therefore it cannot be used directly to quantify new samples. In this sense, calibration techniques are required for quantification.
Different univariate and multivariate calibration techniques have been tested and compared. Univariate techniques correspond to peak area and peak height calibration. Multivariate techniques correspond to PLS and poly-PLS.

5.3.3.1 Univariate calibration

Peak area calibrations have been performed adding up the areas of each one of all the peaks related to the substance and then fitting a polynomial of a certain order to the relationship between substance concentration and peaks area in order to build the calibration model. In this work, we define the peak area as the integral of the peak corresponding to the FWHM (Full Width Half Maximum) region, that is, the sum of intensities above the 50% of the maximum. The polynomial order is optimized using the cross-validation procedure described in section 5.3.3.3. Peak height calibrations have been performed in a similar way to area calibration but taking the maximum of the monomer peak.

5.3.3.2 Partial least squares (PLS) and nonlinear polynomial PLS (poly-PLS)

In the present work, PLS and poly-PLS methods have been applied to the original matrices of spectral responses. In these cases, the number of latent variables (PLS and poly-PLS) and the polynomial order (poly-PLS) were optimized using a cross validation procedure in order to build the calibration model (see section 5.3.3.3).

The main proposal in this work is to build multivariate calibration models using the concentration profiles extracted by a pre-processing MCR-ALS step. For that purpose, PLS and poly-PLS have also been applied to the matrices constructed using the concentration profiles of monomer and dimer (obtained from MCR-ALS). In these cases, the number of latent variables was set to be 2 (monomer and dimer) and only the polynomial order (poly-PLS) was needed to be optimized using the cross validation procedure (see section 5.3.3.3).

5.3.3.3 Cross validation: Leave-One-Block-Out

Cross-validation can be used to achieve two main objectives: assessing the performance of the different calibration techniques (univariate and multivariate cases) and optimizing some parameters (the number of latent variables in PLS and poly-PLS and the polynomial order in the univariate techniques and the poly-PLS case).

The cross-validation procedure used in the present work corresponds to leave-one-block-out (LOBO). First of all, the set of spectra corresponding to the first and last measured concentrations are always used to build the calibration model, which means that this set of samples is not available for validation. The reason is that we are interested in predicting concentrations within a certain range and not out of this range. This requirement could not be fulfilled if the first or last concentration were taken out to be validated. Secondly, leaving one block out means that, given a substance, the set of spectra corresponding to a certain test concentration is taken out to be validated and the remaining set of samples is used to build the calibration model. In other words, the estimation dataset never has the concentration value that is going to be predicted. In this way, the interpolation performance of the model is tested. The set of scans to be validated for each concentration is used to calculate the root mean squared error of validation (RMSEV). This procedure is repeated as many times as concentrations to be validated are (Table 5.1 shows the measured concentrations per each substance). Each validated concentration has its own RMSEV, therefore an averaged RMSEV can be calculated giving the final root mean squared error of cross-validation (RMSECV). See section 5.3.4 to see how the RMSEV and RMSECV are computed.

The RMSECV can be calculated for a different number of latent variables and for a different number of polynomial orders. For the univariate techniques (area and peak height), the polynomial order which minimizes the RMSECV is taken to build the calibration model. For the PLS case, the number of latent variables which minimizes the RMSECV is considered optimum and is taken to build the calibration model. For the poly-PLS case, the combination of the number of latent variables and polynomial order which minimizes the RMSECV is taken to build the poly-PLS calibration model.

In the proposed method, when MCR-ALS is applied as prior step to multivariate calibration, only the polynomial order needs to be optimized in the poly-PLS case, since the number of latent variables is set to be 2 (monomer and dimer) by MCR-ALS.

5.3.4 Figures of merit

In order to assess the performance of MCR-ALS, the explained variance is used as figure of merit:

(Eq. 5.1)
$$EV(\%) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(C \cdot S^T)^2_{m,n}}{\sum_{m=1}^{M}\sum_{n=1}^{N}D_{m,n}} \cdot 100$$

where *C* and *S* correspond to the concentration profiles and spectra matrices recovered by MCR-ALS algorithm and *D* corresponds to the original matrix of spectral responses. The explained variance is equivalent to the recovered power in (Eq. 4.9).

In the case of calibration techniques (univariate and multivariate), it is important to assess the error of prediction in the quantification of new samples. To achieve this, the root mean squared error of validation is calculated:

(Eq. 5.2)
$$RMSEV = \sqrt{\frac{\sum_{v=1}^{V}(c_v - \hat{c}_v)^2}{V}}$$

where c_v corresponds to the original concentration, $\hat{c}_v$ to the concentration predicted by the calibration model, and *V* corresponds to the number of validated samples.

The set of spectra to be validated for each concentration is used to calculate a RMSEV (Eq. 5.2). Therefore, for each concentration a RMSEV is calculated. In the end, an averaged RMSEV can be obtained:

$$RMSECV = \frac{\sum\limits_{i=1}^{I} RMSEV_i}{I}$$

(Eq. 5.3)

where *RMSECV* is the root mean squared error of cross-validation, *RMSEV$_i$* corresponds to the validation error calculated using (Eq. 5.2) for a particular concentration, and *I* corresponds to the number of validated concentrations. This result is presented as a percentage of the maximum substance concentration (see Table 5.2).

The squared correlation coefficient *R^2* also gives a measure of the quality of the prediction. It assesses the correlation between the predicted concentrations by the calibration model and the expected concentrations (Table 5.1). The quality of the prediction is better as *R^2* is closer to 1 ($0 \leq R^2 \leq 1$).

5.3.5 Algorithms implementation details

All calculations were performed using MATLAB 7.9 (MathWorks, Massachusetts USA). MCR-ALS codes are available on the Internet (Tauler & Juan 2010). MATLAB routines for SIMPLISMA (function *purity*), PLS (function *pls*) and poly-PLS (function *polypls*) are available in PLS toolbox 5.8 by Eigenvector Research.

5.4 Results and discussion

Measured IMS spectra for 2-butanone and ethanol include a protonated monomer (Eq. 1.5) and a proton-bound dimer (Eq. 1.6). 2-butanone spectra at different concentrations after preprocessing are shown in Figure 5.2(a). The spectra show four main peaks. The first peak with a reduced mobility K_0=2.10 cm^2/(V·s) is related to the ionization of water vapour in the ambient air within the reactant region and is always present. The second peak with K_0=1.95 cm^2/(V·s) is related to the protonated monomer of 2-butanone. The proton-bound dimer of the analyte appears at high concentrations with K_0=1.64 cm^2/(V·s), and an additional third peak, whose behaviour is strongly correlated with proton-bound dimer, appears at the right of proton-bound dimer with K_0=1.55 cm^2/(V·s).

Figure 5.2(b) represents the ethanol spectra at different concentrations. In this case, the spectra show five peaks. The first two peaks (RIP0 and RIP1) are related to reactant ions from the Ni^{63} ionization source with K_0=2.35 cm^2/(V·s) and K_0=2.10 cm^2/(V·s) respectively. RIP0 is related to the ionization of ammonia which is present in small

concentrations in the ambient air within the instrument and is very difficult to remove it from the spectrometer. RIP1 is related to the ionization of water vapour. Protonated monomer comes out with K_0=1.99 cm^2/(V·s), and the proton-bound dimer shows up at K_0=1.83 cm^2/(V·s). The last peak appears with K_0=1.65 cm^2/(V·s) has a behaviour which is strongly correlated with ethanol dimer.
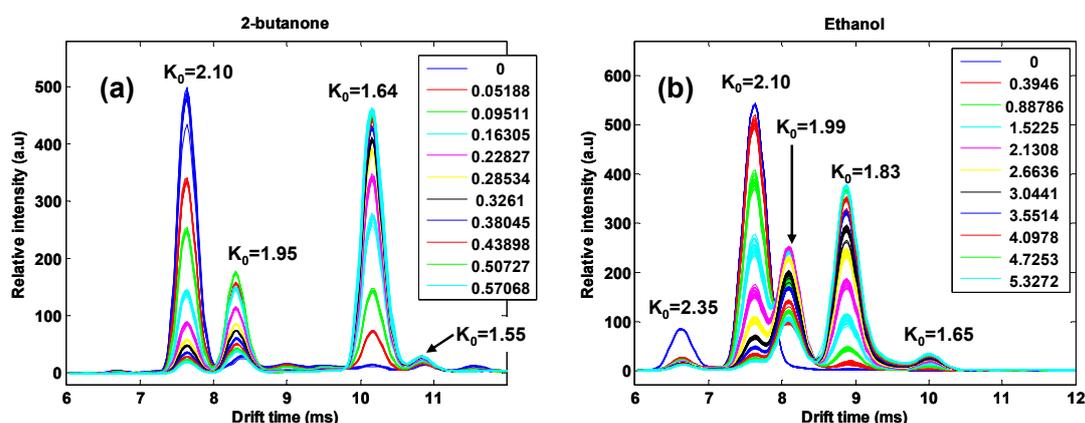
Figure 5.2. IMS spectra for 2-butanone and ethanol at different concentrations (legends in ppm units). a) 2-Butanone. b) Ethanol.

To the best of our knowledge, only in the case of ethanol, K_0 values have been reported using Ni[63] ionization sources, but only regarding the monomer peak (1.91 cm^2/(V·s) (Eiceman & Karpas 2005) and 1.99 cm^2/(V·s) (Sielemann et al. 2001)). This is in agreement with our experiments (1.99 cm^2/(V·s)). Other works reported K_0 values (monomer and dimer) for ethanol using other ionization sources, therefore these values cannot be compared with those obtained in our experiments. For instance, using a 10.6eV UV lamp (Sielemann et al. 2001) or a Tritium ionization source (Tiebe et al. 2009).

In the case of 2-butanone, only K_0 values for the monomer peak have been reported and none of these values have been obtained using a Ni[63] ionization source. For instance, using a UV hydrogen lamp (Leasure et al. 1986) or a high speed capillary column (HSCC-UV-IMS) (Xie et al. 2002).

MCR-ALS was applied to both datasets to resolve the evolution of formed species of 2-butanone and ethanol. SIMPLISMA was used to extract initial estimations for spectra and concentration profiles prior to MCR-ALS. Non-negativity, unimodality and closure were the constraints used within the ALS loop. The analysis was performed using the spectra region from 6ms to 12ms where relevant peaks appear. SIMPLISMA was applied imposing 3 components in the case for 2-butanone and 4 components in the case for ethanol.
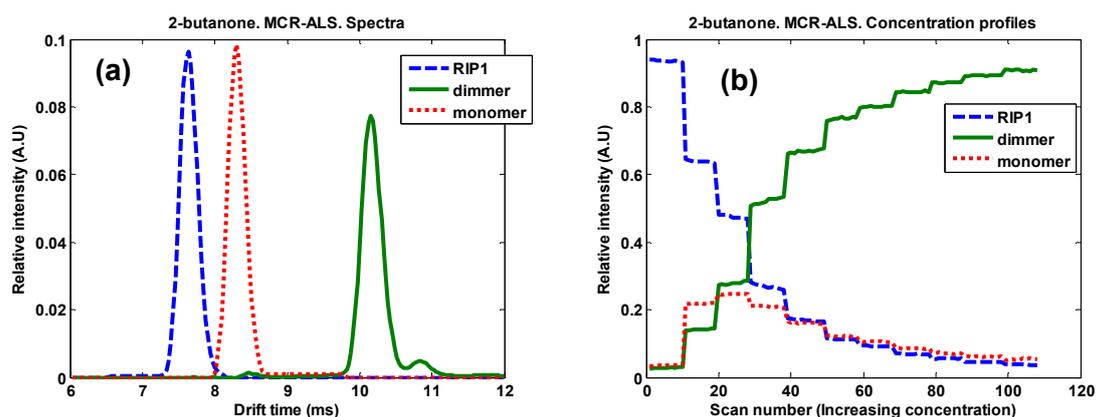
Figure 5.3. 2-butanone. (a) Recovered spectra by MCR-ALS. (b) Recovered concentration profiles by MCR-ALS.
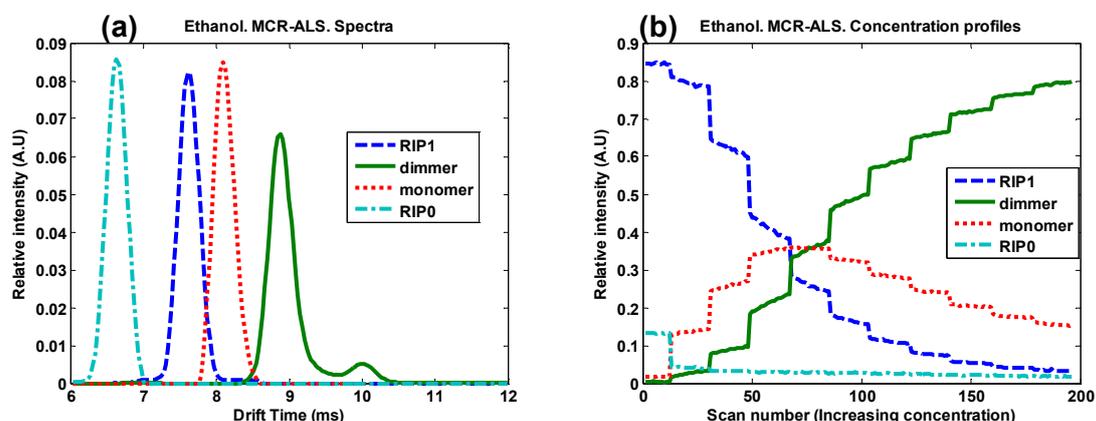
Figure 5.4. Ethanol. (a) Recovered spectra by MCR-ALS. (b) Recovered concentration profiles by MCR-ALS.

Figure 5.3 and Figure 5.4 show the results by MCR-ALS for 2-butanone and ethanol respectively. 2-butanone and ethanol have a nonlinear behaviour as substance concentration increases. Moreover, since protonated-bound dimers appear at high concentrations and their behaviour differs from that of the monomers, the concentration of the substances needs to be explained using more than one component in SIMPLISMA and MCR-ALS. In the studied cases, MCR-ALS is able to resolve the different components properly. As it can be seen in Figure 5.3(b) and Figure 5.4(b), the intensity of the reactant ion peaks decreases as substance concentration increases. Although protonated monomers start increasing their intensities at low concentrations, they reach their maximum intensity in a certain substance concentration and then start to drop off. At the same time, proton-bound dimer peaks increase their intensity when substance concentration rises further, but

they reach a saturated behaviour at very high concentrations. Furthermore, sometimes clustering formation takes place between the analyte and water molecules either in the reactant region or in the drift tube (Eiceman & Karpas 2005), as a result of this chemical process a new peak could appear in the signal. This is observed in the MCR-ALS results where a secondary peak appears in the dimer components; the peak located at the right of the dimers is related to a product formed by the proton-bound dimer and a water molecule.

The explained variance (Eq. 5.1) by MCR-ALS models were 99.7% for 2-butanone and 99.8% for ethanol, which indicates that MCR-ALS models are able to explain almost the total variance of the original data matrix and at the same time provide an easy interpretation for the different contributions.

Once we have modelled the qualitative evolution of monomers and dimers for 2-butanone and ethanol using MCR-ALS, this information can be seized in order to build calibration models for quantification. Since substances present a strong nonlinear behaviour as concentration increases and especially monomers and dimers peaks, polynomial PLS should be used to construct the calibration model rather than PLS or any other univariate technique. The output information from MCR-ALS concentration profiles is used to construct a new matrix (X) with dimensions M x N, where M is the total number of samples (including all scans at all concentrations) and $N = 2$ (monomer and dimer concentration profiles from MCR-ALS). A matrix of concentrations (Y) with dimensions M x R can also be constructed, where $R=1$ since we only have one substance per model. Using X and Y matrices (Eq. 1.13), an optimum calibration model can be built using the cross-validation methodology explained in section 5.3.3.3. Moreover, this cross-validation methodology can be used to assess the performance of the calibration model.
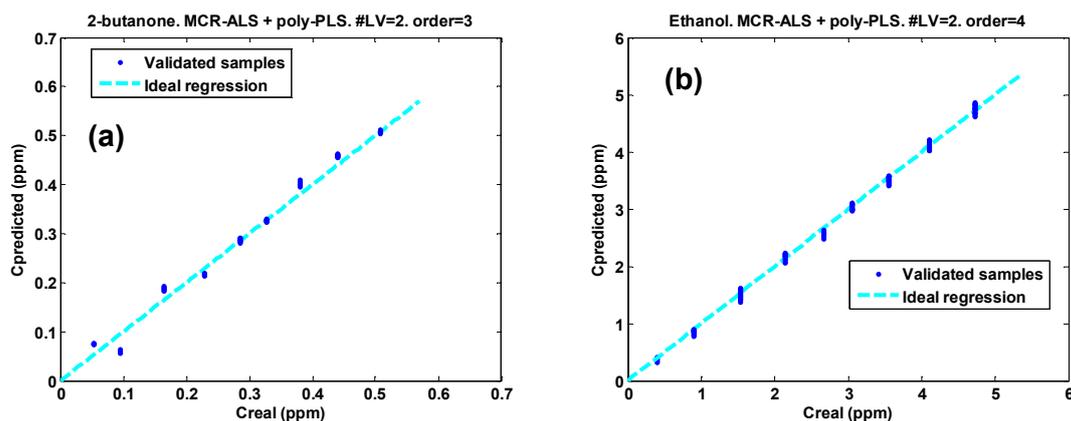
Figure 5.5. Predicted concentrations vs. substance concentrations for validation samples projected over constructed poly-PLS models. (a) Predicted 2-butanone concentrations using poly-PLS models with 2 latent variables and polynomial order = 3. (b) Predicted ethanol concentrations using poly-PLS models with 2 latent variables and polynomial order = 4.

Figure 5.5 shows the predicted concentrations versus the original concentrations for 2-butanone and ethanol using poly-PLS as calibration method after obtaining MCR-ALS concentration profiles. The figure shows only the validation results. The reader should notice that although the validation results are depicted on the same graph, each set of scans (belonging to a particular substance concentration) has a different calibration model (built from leave-one-block-out cross validation method). The optimum polynomial order was found to be 3 for 2-butanone data and 4 for ethanol data. The RMSECV (Eq. 5.3) was 5.6% (relative to full scale input range) for 2-butanone and 1.2% for ethanol (relative to full scale input range). The squared correlation coefficient was 0.98 for 2-butanone and 0.998 for ethanol. The results show that prediction accuracy is quite good using the proposed method.

In order to compare with the proposed method, univariate and multivariate calibration models can also be built without using MCR-ALS concentration profiles. Figure 5.6 and Figure 5.7 show predicted concentrations for 2-butanone and ethanol respectively, using the same cross-validation methodology.
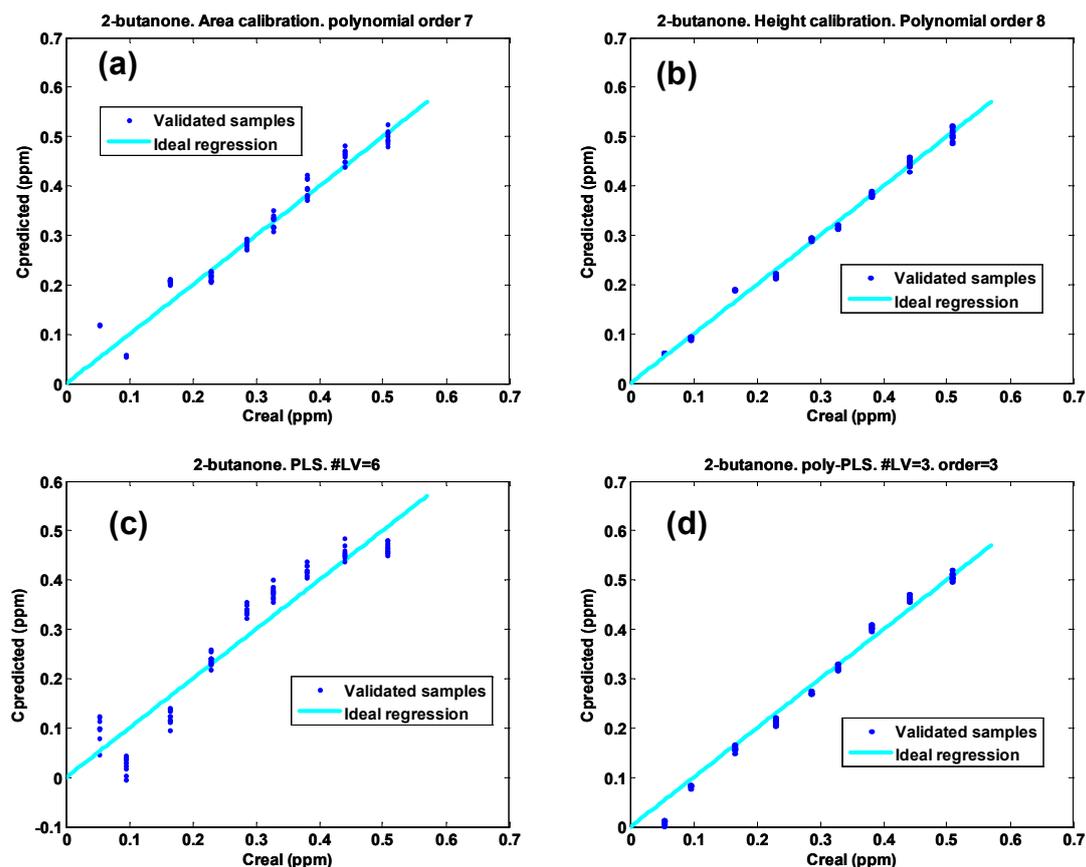
Figure 5.6. Predicted concentrations vs. substance concentrations for validation samples projected over different calibration models. (a) Predicted 2-butanone concentrations using area calibration and fitting a polynomial of 7th order. (b) Predicted 2-butanone concentrations using height calibration and fitting a polynomial of 8th order. (c) Predicted 2-butanone concentrations using PLS models with 6 latent variables. (d) Predicted 2-butanone concentrations using poly-PLS models with 3 latent variables and polynomial of order 3.
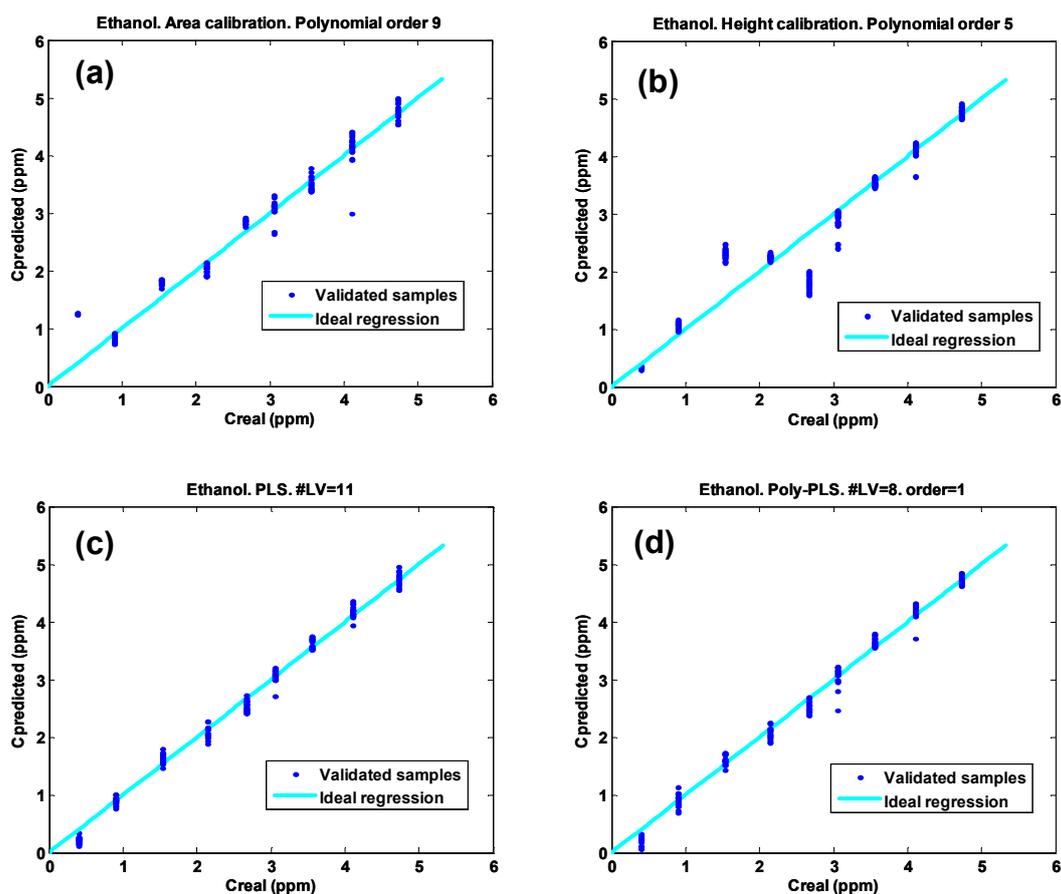
Figure 5.7. Predicted concentrations vs. substance concentrations for validation samples projected over different calibration models. (a) Predicted ethanol concentrations using area calibration and fitting a polynomial of 9th order. (b) Predicted ethanol concentrations using height calibration and fitting a polynomial of 5th order. (c) Predicted ethanol concentrations using PLS models with 11 latent variables. (d) Predicted ethanol concentrations using poly-PLS models with 8 latent variables and polynomial of order 1.

Previous results were obtained using spectra acquired within one day (day 1). Numerical results comparing univariate and multivariate techniques using and not using MCR-ALS concentration profiles are presented in Table 5.2. Additionally, Table 5.2 presents results from replicated datasets obtained in different days. Henceforth the analysis of the results is referred to day 1, but similar analyses could be done from day 2 and 3.

Calibration method	R^2		RMSECV (% max conc)	
	2-butanone	Ethanol	2-butanone	Ethanol
Peak area calibration (U)				
Day 1	0.96	0.95	5.6	6.3
Day 2	0.998	0.96	1.1	5.8
Day 3	0.87	0.96	12	6.1
Peak height calibration (U)				
Day 1	0.993	0.91	2.3	7.9
Day 2	0.94	0.90	6.4	8.4
Day 3	0.990	0.92	2.7	7.8
PLS (M)				
Day 1	0.91	0.993	7.7	2.3
Day 2	0.998	0.995	1.2	1.8
Day 3	0.996	0.992	1.4	2.3
poly-PLS (M)				
Day 1	0.992	0.991	3.0	2.6
Day 2	0.998	0.992	2.7	2.3
Day 3	0.995	0.992	1.6	2.4
MCR-ALS + PLS (M)				
Day 1	0.85	0.97	10	5.7
Day 2	0.994	0.97	2.4	5.2
Day 3	0.89	0.97	8.8	5.8
MCR-ALS + poly-PLS (M)				
Day 1	0.98	0.998	5.6	1.2
Day 2	0.9990	0.998	0.88	1.2
Day 3	0.997	0.998	1.2	1.3

Table 5.2. Comparison between different optimized calibration methods using leave-one-block-out cross validation with datasets obtained in different days. Results include univariate (U) and multivariate (M) methods. The best results for each day are shown shaded.

As it can be seen in Figure 5.6 and in Table 5.2, univariate techniques can provide good results if peaks in the spectra do not appear overlapped and thus can be easily identified to calculate their area or extract their height, this is the case for 2-butanone. However, when peaks appear overlapped (case of ethanol) these techniques can fail since contributions from other peaks appear in the region of the peak of interest (Figure 5.7). In situations with a high overlap between peaks the use of univariate calibration techniques can be unfeasible, unless a prior deconvolution step is carried out. Using multivariate techniques better calibration models than univariate techniques can be built as it was already proved by Fraga et. al in (Fraga, Kerr & Atkinson 2009).

Regarding multivariate techniques, although PLS and poly-PLS directly applied to IMS spectra provide similar prediction accuracy to that of the methodology proposed in the present work, the optimum number of latent variables is too high. For the PLS case, 6 and 11 latent variables for 2-butanone and ethanol respectively. For the poly-PLS case, 3 and 8 latent variables for 2-butanone and ethanol respectively. This fact hinders the qualitative interpretation of the results since many different contributions need to be taken into account in order to understand the chemical process involved in the substance behaviour as concentration increases. Moreover, since no constraints are imposed to the regression coefficients, negative values which do not have any physical and chemical meaning can be found.

For instance, Figure 5.8 shows the scores and loadings from a poly-PLS calibration model with the same number of latent variables as the number of components used in MCR-ALS for 2-butanone (Figure 5.8(a) and (b)) and ethanol (Figure 5.8(c) and (d)). The cross-validation procedure has been applied in order to optimize the polynomial order. The difficulty to interpret the results compared to MCR-ALS solutions (Figure 5.3 and Figure 5.4) is shown since many contributions need to be taken into account. If the optimum calibration model includes more latent variables, although prediction can be better, the interpretation of the results is even more difficult, which is the case for the results presented in Table 5.2 (day 1).
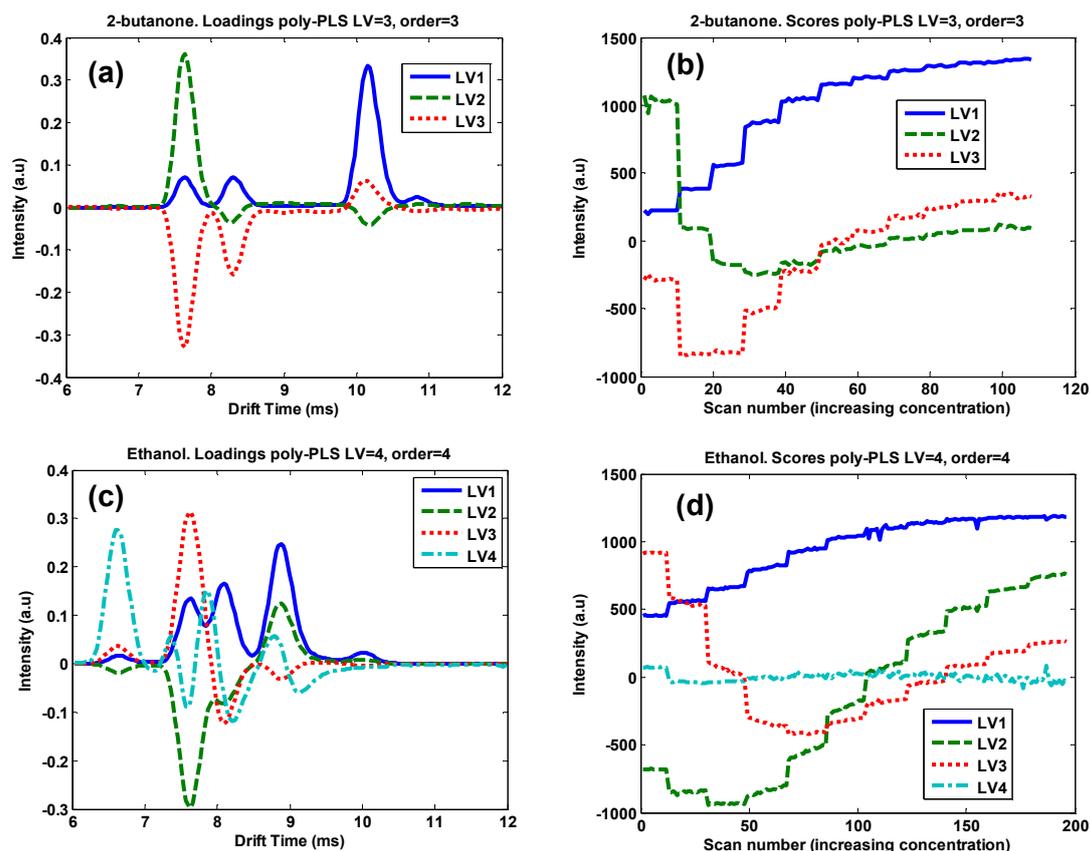
Figure 5.8. Scores and loadings from poly-PLS calibration models using the same number of latent variables as the number of components used to build MCR-ALS models. (a) Loadings for 2-butanone. (b) Scores for 2-butanone. (c) Loadings for ethanol. (d) Scores for ethanol.

Results in Table 5.2 and Figure 5.5 show how the methodology proposed in the present work provides similar or better prediction accuracy compared to other univariate or multivariate methodologies. Moreover, the number of parameters to be optimized by cross-validation in PLS and poly-PLS is one less since the number of latent variables is set by the number of components used in MCR-ALS, thus reducing the complexity of the calibration model. Furthermore, interpretation of the results is much easier since it can be done directly analyzing MCR-ALS concentration profiles.

Similar results were obtained from different replicates in different days, with the calibration and validation samples obtained within the same day. Otherwise, instrumental drift would degrade prediction accuracy. This point has been already observed by different authors, see e.g. Fraga and references there in (Fraga, Kerr & Atkinson 2009).

From days 2 and 3, using MCR-ALS as prior step to poly-PLS calibration provides the best results. PLS and poly-PLS directly applied to IMS spectra also provide good prediction accuracy but, as explained before, the interpretation of the chemical process is difficult since the optimum number of latent variables is too high.

As it is shown in Table 5.2, calibration models can be built within the same day and be used for prediction within the same day; however, there is a large variability in the evolution of monomer and dimer among different days, especially in the 2-butanone case. This variability is much less in the case of ethanol. This result suggests that in general for some substances, calibration models built in one day cannot be used to predict new samples measured in another day. In the ethanol case, although calibration models built in one day could be used in different days, the study of the use of calibration models for prediction in different days is out of the scope of the present thesis.

5.5 Conclusions

This work presents a methodology to be applied to IMS spectra which combines the advantages of MCR-ALS for qualitative interpretation and poly-PLS for quantitative prediction of new samples which present a strong nonlinear behaviour as substance concentration increases.

MCR-ALS has been demonstrated to be a suitable method to the study of ion mobility second order data. Using SIMPLISMA and MCR-ALS, IMS spectra are resolved into pure components and a qualitative estimation for the spectral and concentration profiles of these components is obtained. MCR-ALS allows the description of the chemical changes observed when concentration increases. Although protonated monomer and proton-bound dimer belong to the same substance, their IMS peaks were modelled as separate components in the studied analytes (2-butanone and ethanol). Therefore, MCR-ALS provides a powerful way to obtain a physical and chemical interpretation of the system even in the presence of strong nonlinear behaviours. Although this qualitative information cannot be used directly to perform a calibration, it can be used effectively by a multivariate calibration technique such as poly-PLS to build calibration models. In this work, it has been demonstrated that these models provide similar or better predictive capacity compared to standard univariate and multivariate methodologies.

For the studied datasets, quantitative results show how standard multivariate calibration techniques work in general better than univariate techniques, especially when peaks in the spectra appear overlapped. Multivariate techniques are able to model nonlinear behaviours adding more components to the model. The datasets included strong nonlinear behaviours as substances concentration increased. While PLS is able to handle slightly nonlinear behaviours, strong nonlinear evolutions are better modelled using poly-PLS. Although prediction accuracy is similar, the results obtained from these standard techniques are often difficult to interpret, since, in order

to model nonlinearities, the number of latent variables in the model is usually higher than the number of peaks. Using MCR-ALS prior to the calibration step provides a way to interpret properly the results and fix the number of latent variables, thus reducing the complexity of the calibration model. Moreover, since the number of latent variables is set to be equal to the number of pure components in MCR-ALS, the number of parameters to be optimized by cross-validation is one less.

Additionally, results obtained from different replicates have been presented. The results suggest that calibration models can be used for prediction in the same day but not between days. This is especially significant in the case of 2-butanone. Using an IMS calibration model built in one day to predict new samples in another day is an interesting topic for future work.

5.6 References

Antunes, MC, J. Simao, JE, Duarte, AC & Tauler, R 2002, 'Multivariate curve resolution of overlapping voltammetric peaks: quantitative analysis of binary and quaternary metal mixtures', *Analyst*, vol. 127, no. 6, pp. 809-817.

Armenta, S & Blanco, M 2012, 'Ion Mobility Spectrometry: A Comprehensive and Versatile Tool for Occupational Pharmaceutical Exposure Assessment', *Analytical Chemistry*, vol. 84, no. 10, pp. 4560-4568.

Azzouz, T & Tauler, R 2008, 'Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples', *Talanta*, vol. 74, no. 5, pp. 1201-1210.

Baffi, G, Martin, EB & Morris, AJ 1999, 'Non-linear projection to latent structures revisited: the quadratic PLS algorithm', *Computers & Chemical Engineering*, vol. 23, no. 3, pp. 395-411.

Bro, R & Jong, Sd 1997, 'A fast non-negativity-constrained least squares algorithm', *Journal of Chemometrics*, vol. 11, pp. 393-401.

Eiceman, GA & Karpas, Z 2005, *Ion Mobility Spectrometry*, second edn, CRC Taylor and Francis Group, Boca Raton.

Fraga, CG, Kerr, DR & Atkinson, DA 2009, 'Improved quantitative analysis of ion mobility spectrometry by chemometric multivariate calibration', *Analyst*, vol. 134, no. 11, pp. 2329-2337.

Harrington, PDB & Chen, P 2004, 'Equilibrium Modeling of Ion Mobility Spectra', in *The conference Proceedings of the Thirteenth International Workshop on IMS*, pp. 160-181.

Leasure, CS, Fleischer, ME, Anderson, GK & Eiceman, GA 1986, 'Photoionization in air with ion mobility spectrometry using a hydrogen discharge lamp', *Analytical Chemistry*, vol. 58, no. 11, pp. 2142-2147.

References

Lu, Y, O'Donnell, RM & Harrington, PB 2009, 'Detection of cocaine and its metabolites in urine using solid phase extraction-ion mobility spectrometry with alternating least squares', *Forensic Science International*, vol. 189, no. 1-3, pp. 54-59.

Ochoa, ML & Harrington, PB 2005, 'Chemometric studies for the characterization and differentiation of microorganisms using in situ derivatization and thermal desorption ion mobility spectrometry', *Analytical Chemistry*, vol. 77, no. 3, pp. 854-863.

Pomareda, V, Calvo, D, Pardo, A & Marco, S 2010, 'Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra', *Chemometrics and Intelligent Laboratory Systems*, vol. 104, no. 2, pp. 318-332.

Puton, J, Holopainen, SI, Makinen, MA & Sillanpaa, MET 2012, 'Quantitative Response of IMS Detector for Mixtures Containing Two Active Components', *Analytical Chemistry*, vol. 84, no. 21, pp. 9131-9138.

Rosipal, R 2008, 'Nonlinear Partial Least Squares An Overview', in *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pp. 169-189.

Savitzky, A & Golay, MJE 1964, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', *Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639.

Sielemann, S, Baumbach, JI, Schmidt, H & Pilzecker, P 2001, 'Detection of alcohols using UV-ion mobility spetrometers', *Analytica Chimica Acta*, vol. 431, no. 2, pp. 293-301.

Tauler, R & Juan, Ad 2010, *http://www.mcrals.info/* in Multivariate Curve Resolution Homepage.

Tauler, R, Kowalski, B & Fleming, S 1993, 'Multivariate curve resolution applied to spectral data from multiple runs of an industrial process', *Analytical Chemistry*, vol. 65, no. 15, pp. 2040-2047.

Tiebe, C, Miessner, H, Koch, B & Hübert, T 2009, 'Detection of microbial volatile organic compounds (MVOCs) by ion-mobility spectrometry', *Analytical and Bioanalytical Chemistry*, vol. 395, no. 7, pp. 2313-2323.

Windig, W, Gallagher, NB, Shaver, JM & Wise, BM 2005, 'A new approach for interactive self-modeling mixture analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, pp. 85-96.

Wold, S 1992, 'Nonlinear partial least squares modelling II. Spline inner relation', *Chemometrics and Intelligent Laboratory Systems*, vol. 14, no. 1-3, pp. 71-84.

Wold, S, Kettaneh-Wold, N & Skagerberg, B 1989, 'Nonlinear PLS modeling', *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 53-65.

Xie, ZX, Sielemann, SS, Schmidt, HS, Li, FL & Baumbach, JIB 2002, 'Determination of acetone, 2-butanone, diethyl ketone and BTX using HSCC-UV-IMS', *Analytical and Bioanalytical Chemistry*, vol. 372, no. 5, pp. 606-610.

Yang, H, Griffiths, PR & Tate, JD 2003, 'Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra', *Analytica Chimica Acta*, vol. 489, no. 2, pp. 125-136.

Zamora, D & Blanco, M 2012, 'Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration', *Analytica Chimica Acta*, vol. 726, no. 0, pp. 50-56.

Zheng, P, Harrington, PdB & Davis, DM 1996, 'Quantitative analysis of volatile organic compounds using ion mobility spectrometry and cascade correlation neural networks', *Chemometrics and Intelligent Laboratory Systems*, vol. 33, pp. 121-132.

Chapter 6

Chemical Plume Source Localization with Multiple Mobile Partially Selective Sensors using Bayesian Inference in the presence of Chemical Background Interference

6.1 Introduction

Navigation experiments towards chemical sources are strongly limited by the rapid decay in the chemical concentration of the source of interest with increasing distance. Poor detection limits result in a reduced area where the plume can effectively be detected. This is especially important in applications where the search zone has an area of hundreds of thousands of square meters. Unless the source is very strong, the task of source localization becomes practically impossible. In such situations, it becomes very important to set the detection thresholds very close to the noise level, but this would result in a high number of false alarms and most localization algorithms would fail catastrophically. To the best of our knowledge, none of the published methods have addressed this problem.

On the other hand, low cost chemical sensors and even medium-priced detectors (e.g. ion mobility spectrometers) have only limited selectivity. In any real scenario, there could be background levels of a multitude of chemicals due to environmental pollution. Due to limited selectivity, there will be substances which will produce interference in the detector reading, hindering the detection and localization tasks. The combination of some detector electronic noise and mainly interfering chemical agents result in background readings that may change with time and the position of the detector. These shifts in background levels hamper the selection of an optimum threshold. As far as we know, this problem has not been tackled previously in the source localization literature.

In order to address these problems, probabilistic approaches like plume mapping Bayesian methods appear to be a good choice. Such techniques have been effective in many areas of robotics and provide a good starting point to tackle the problem. Pang and Farrell published a source-likelihood mapping approach based on Bayesian

inference in 2006 (Pang & Farrell 2006). The main idea behind the algorithm consists in implementing a stochastic approach (Farrell et al. 2002; Papoulis 1984b) for plume modelling and in estimating the most likely source position taking into account the sequence of detection/non-detection events and fluid flow measurements along the robot's trajectory. Although in the original work, results on plume tracking are presented, it is not strictly necessary to track the plume to obtain a good estimation of the source location. Recursively building a probability map using Bayesian inference, the most likely source position is estimated during the robot's mission.

However, this aforementioned algorithm uses only binary detection events. In other words, it does not use the chemical concentration information to build the probability map, since it only considers the concentrations above a certain threshold as detection or non-detection events. Moreover, after setting the threshold level, the approach assumes that the rate of false alarms is very low; such an assumption is far from the truth in a real scenario where background signals arise unless the threshold is set at a high level– but using this option seriously reduces the maximum plume detection distance. Therefore, there is a trade-off; on the one hand, the threshold needs to be set low enough (close to the sensor detection limit) if chemicals from the source are to be detected at large distances; on the other hand, it needs to be high enough to prevent false alarms. So, how to set the threshold level becomes a critical issue in real environments using existing approaches, especially when no information about the background or the source strength is available.

One of the motivations for the present thesis was to improve the Bayesian plume source localization algorithm, previously described by Pang and Farrell (Pang & Farrell 2006), using the chemical concentration and extending it to real environments where background signals may arise. In our proposal, the algorithm assesses the posterior probability that a given chemical concentration reading comes from the background or from a source emitting at a greater distance with a specific release rate. This removes the need to set any concentration threshold. Thus, Pang's algorithm is reformulated for use with continuous concentration readings instead of merely binary detections. Moreover, the algorithm is extended to work with multiple mobile robots, simultaneously integrating their readings. This new approach requires a probabilistic model for the background and for the plume which are described in the following sections.

This chapter is structured as follows: in Section 6.2 a summary of the work by Pang and Farrell is given which establishes the basis of our proposal. Section 6.3 explains the different parts of our algorithm. Section 6.4 describes the assumptions and the simulation scenario devised to test both approaches. Results from simulations are

discussed in detail in section 6.5. In order to support the conclusions obtained from simulations, real-world data is needed, thus the experimental set-up built to obtain these data is described in section 6.6 and their results are analyzed in section 6.7. Finally section 6.8 closes the chapter with the conclusions.

6.2 Brief description of Pang and Farrell's work (binary-based approach)

For computational feasibility, as illustrated in Figure 6.1, the search area is divided into rectangular cells C_i, where i is an index running from 1 to N_c, where N_c is the total number of cells in the grid area. Then the area contains m_x x m_y rectangular cells each of size L_x x L_y, where $N_c = m_x$ x m_y and L_x and L_y are the cell lengths in the x and y directions of the grid map, respectively. Let $0 \leq \alpha'_i \leq 1$ represent the probability that there is a chemical source in C_i. It is assumed that the search area contains exactly one source, hence $\sum_{i=1}^{N_c} \alpha'_i = 1$. Initially ($t_0$), if no information about the source location is available, we assume uniform priors, i.e. all cells are initialized to be equally likely to contain the chemical source: $\alpha'_i(t_0) = (1/N_c); \forall i \in [1, N_c]$. If previous information is available, different selections of the priors are possible.
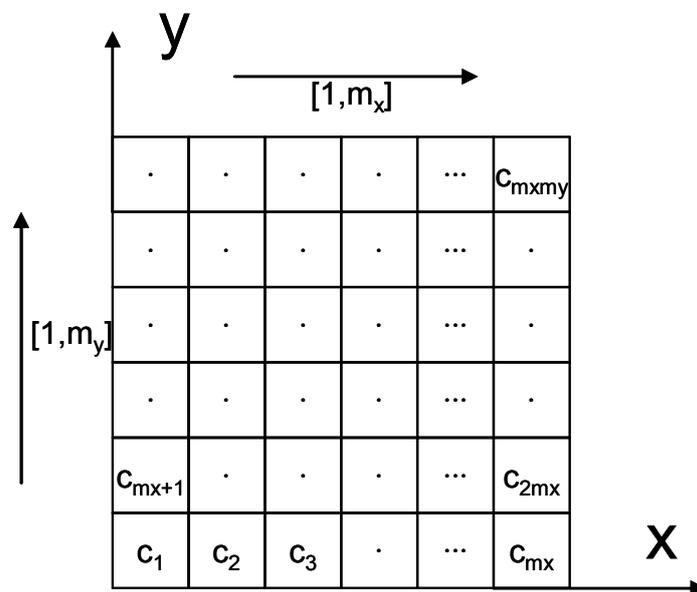
Figure 6.1. Cellular subdivision of the region to be searched.

The source is modelled according to a stochastic plume model based on localized chemical filaments ("puffs") (Farrell et al. 2002). In this model, it is supposed that chemical filaments are continuously released from the source. Although the source is

continuously emitting filaments, let us start by studying the case of a single chemical release. The location of a filament at time t_k released at time t_l ($t_l < t_k$) is modelled as:

(Eq. 6.1)
$$X(t_l, t_k) = \int_{t_l}^{t_k} U(X(t)) \, dt + \int_{t_l}^{t_k} N(t) \cdot dt + X_s$$

where $X(t_l, t_k) = (x, y)$ is the chemical-filament location at time t_k given that the chemical filament was released from a source located at $X_s = (x_s, y_s)$ at time t_l. $U = (u_x, u_y)$ is the mean flow velocity and $N = (n_x, n_y)$ is a Gaussian random process with zero mean and (σ_x^2, σ_y^2) variance. If the released chemical filament at time t_l is detected within cell C_j at time t_k, what is the probability that the source is located in some cell C_i? Solving (Eq. 6.1) for the possible source location yields:

(Eq. 6.2)
$$X_S(t_l, t_k) = X_j - V(t_l, t_k) - W(t_l, t_k)$$

where X_j is the center of cell C_j; $V(t_l, t_k) = \left(v_x(t_l, t_k), v_y(t_l, t_k) \right) = \sum_{i=l}^{k-1} U(X_r(t_i)) \cdot dt$ is the flow vector along the trajectory of the robot, where $X_r(t_i) = (x_r(t_i), y_r(t_i))$ is the position of the sensor at time t_i; and $W(t_l, t_k)$ is a zero-mean Gaussian random variable.

For each single chemical release, the algorithm computes $S_{ij}(t_l, t_k)$ which stands for the probability of there being a source in C_i which released a single filament at time t_l, given that the sensor detected the chemical in C_j at time t_k ($t_k > t_l$). $S_{ij}(t_l, t_k)$ is a function of the relative position of cell C_i and cell C_j, the flow vector $V(t_l, t_k)$, and the time period $[t_l, t_k]$.

If the sensor does not detect any chemical in C_j at time t_k, there is still a certain probability of being a source in C_i. μ is the probability of detecting a chemical, given that there is detectable chemical in the cell C_j. Then, assuming that there is a source in C_i, the probability of detecting the chemical in C_j is $[\mu \cdot S_{ij}(t_l, t_k)]$ and the probability of not detecting the chemical is $[1 - \mu \cdot S_{ij}(t_l, t_k)]$. Detection and nondetection events are decided based on a threshold. $S_{ij}(t_l, t_k)$ is computed for all the cells in the map with i being an index running from 1 to N_c.

This explanation is valid for one single chemical filament, but it is assumed that the source will be continuously emitting chemical filaments, starting at t_0. As the sensor moves and senses the environment, detection and nondetection events are generated along the robot's trajectory. These events are incorporated recursively into the algorithm in order to update the source probability map: $\alpha'_{ij}(t_k) = P(A_i \mid B(t_k))$; where

A_i stands for the event that there is a source in cell C_i, $B(t_k)$ stands for a sequence of detection and nondetection events from t_0 to t_k and $\alpha'_{ij}(t_k)$ stands for the probability of cell C_i of containing a source. Please note that, since the sequence $B(t_k)$ depends on the trajectory of the robot, it depends on the index j (related to the cells visited by the robot) and $\alpha'_{ij}(t_k)$ as well; however this index j is not explicitly used in (Pang & Farrell 2006). Events are incorporated using Bayesian inference and considered to be independent.

Given a detection event, applying the law of total probability $\left(P(Z) = \sum_{l=0}^{k-1} P(Z\,|\,Y_l) \cdot P(Y_l) \right)$ and assuming uniform probability distribution for the time of filament emission, $\beta_{ij}(t_0,t_k)$ is computed. This stands for the probability of there being a source continuously emitting filaments in C_i, given that there is detectable chemical in C_j at time t_k :

(Eq. 6.3)
$$\beta_{ij}(t_0,t_k) = \frac{1}{k}\sum_{l=0}^{k-1} S_{ij}(t_l,t_k)$$

Given a nondetection event, $\gamma_{ij}(t_0,t_k)$ is computed. This stands for the probability of there being a source continuously emitting filaments in C_i, given that no chemical was detected in C_j at time t_k:

(Eq. 6.4)
$$\gamma_{ij}(t_0,t_k) = \prod_{l=0}^{k-1}\left[1 - \mu \cdot S_{ij}(t_l,t_k)\right]$$

The source probability map $\alpha'_{ij}(t_k)$ is updated either by means of (Eq. 6.3) or (Eq. 6.4) and, using Bayesian theory, depending on whether the sensor detects or does not detect a chemical filament in cell C_j at time t_k.

Since the false-alarm rate is considered to be very low, detection events are incorporated taking into account that the detection comes from the source. Given a detection event, the map is updated using (Eq. 6.3):

(Eq. 6.5)
$$\alpha'_{ij}(t_k) = N_C \cdot \alpha'_{ij}(t_{k-1})\beta_{ij}(t_0,t_k)$$

On the other hand, nondetection events are incorporated taking into account that it is still possible that the cell contains the source. Given a nondetection event, the map is updated using (Eq. 6.4):

(Eq. 6.6)
$$\alpha'_{ij}(t_k) = \frac{N_C \cdot \alpha'_{ij}(t_{k-1}) \gamma_{ij}(t_0, t_k)}{\sum_{i=1}^{N_C} \gamma_{ij}(t_0, t_k)}$$

$\beta_{ij}(t_0, t_k)$, $\gamma_{ij}(t_0, t_k)$ and $\alpha'_{ij}(t_k)$ are computed for all the cells in the map being *i* an index from 1 to N_c, and *j* the index related to the cells visited by the robot.

Since t_0 is the time when the source starts emitting chemical filaments and t_k is the current measurement time which grows as the exploration time increases, for computational feasibility a certain time-window $(t_k - t_l)$ is considered, and only the last emitted filaments are taken into account. Moreover, only those filaments within the search area are considered in the calculations. This imposes some constraints on the computation of $V(t_l, t_k)$ and $S_{ij}(t_l, t_k)$ and consequently on the calculation of equations (Eq. 6.3) to (Eq. 6.6). The reader is referred to the original work of Pang et. al. for the details (Pang & Farrell 2006).

Despite the fact that Pang's algorithm has been tested successfully using an autonomous underwater vehicle (Farrell, Pang & Li 2005; Pang & Farrell 2006), the work does not address the problem of the presence of interfering substances in the background. Additionally, although all the work is described considering only one robot, the algorithm can easily be extended to work with multiple robots. Independently of the number of robots, for the algorithm the only information needed is: the position where the measurement was obtained, whether it was a detection or a nondetection event, and the fluid flow measurement. However, a new issue needs to be dealt with when using multiple robots. As reported in other works (Li et al. 2011), the binary-based approach is very sensitive to false alarms and unexpected measurements caused by airflow turbulence. Since the sensors mounted on the robots are deployed over the area at different locations and the algorithm uses their readings to build a single probability map, the most likely source location might change too quickly from cell to cell, depending on the location of the last measurement. This effect is worsened if the threshold is too low, since many of readings are considered as detection events. In order to minimize this problem, a filtering step with time constant τ is added to the original algorithm after equations (Eq. 6.5) and (Eq. 6.6). The filtered source probability map is computed as:

(Eq. 6.7)
$$\alpha_{ij}(t_k) = \frac{\alpha_{ij}(t_{k-1}) \cdot (\tau - 1) + \alpha'_{ij}(t_k)}{\tau}$$

where $\alpha'_{ij}(t_k)$ comes from (Eq. 6.5) or (Eq. 6.6); and τ is a parameter which weights current and previous values ($\tau \geq 1$).

6.3 Concentration-based approach

Unlike Pang's work, our proposal is not based on detection and nondetection events. Instead, our approach uses continuous concentration information in order to recursively update the probability map. Therefore, from the point of view of Pang's work, now we always have detection events (concentrations). This new approach requires making some assumptions about the dispersion of the plume (the stochastic model in section 6.3.1) and a background model (section 6.3.2) for each cell of the search area. Given an instantaneously measured concentration c, it is assumed that there are two additive contributions: one due to the background (c_b) and one due to the plume (c_p), thus:

(Eq. 6.8) $$c = c_b + c_p$$

The aim of the algorithm is to estimate whether the concentration c comes from the background at the current cell C_j, or from a source emitting chemicals further away in cell C_i, –it then calculates the probability of having a source at C_i by weighting both possibilities. Since the algorithm is based on continuous readings, no threshold is needed, which eliminates the problem of losing information when concentrations are close to the sensor detection limit.

6.3.1 Stochastic time-averaged plume model

The basis of our stochastic model (Papoulis 1984b) for the chemical plume is the analytical Gaussian plume model (GPM) (Turner 1994; Beychok 2005) which has been pointed in the introduction (section 1.5.1). Previous work has demonstrated that the time-average plume concentration follows a Gaussian distribution along the flow direction (Sutton 1947; Fackrell & Robins 1982; Crimaldi, Wiley & Koseff 2002; Webster & Weissburg 2001). This model has been widely used for its simplicity and is appropriate when dispersion is governed by atmospheric turbulence. The basic expression for the GPM for a continuous release is given in (Eq. 6.9).

(Eq. 6.9) $$C(x,y,z) = \frac{q}{2\pi U_a \sigma_y \sigma_z} \exp\left(-\frac{y^2}{2\sigma_y^2}\right) \exp\left(-\frac{(h-z)^2}{2\sigma_z^2}\right)$$

where *C* is the mean concentration in g/m^3 in a location with coordinates *x* (downwind), *y* (crosswind) and *z* (vertical); *q* is the continuous source release rate in g/s; U_a is the mean wind speed in the downwind direction in m/s; *h* is the plume height in m; and σ_y, σ_z are the diffusion coefficients (in metres) modelled as: $\sigma_y=a{\cdot}x^b$, $\sigma_z=c{\cdot}x^d$, where *a*, *b*, *c* and *d* are parameters obtained from a table (Bakkum & Duijm 2005) and their values depend on the atmospheric conditions which can be organized in six levels (from A-very unstable to F-very stable). In (Eq. 6.9), the resulting concentration distribution is due to the transport of chemicals by advection (due to the mean wind speed) and due to concentration gradients within the plume width (lateral dispersion due to diffusion, but also turbulent mixing).

The decay of mean concentration is exponential, thus concentration levels below the sensor detection limit are very quickly achieved. This issue would make the setting of the threshold level critical; especially, if the source has to be detected far from the release point. The GPM takes into account the time-averaged characteristics of a plume dispersed in a turbulent flow, but the sensors will be responding to the instantaneous plume characteristics. We propose an alternative approach to that published in (Farrell et al. 2002), for which an additional component needs to be added to the GPM in order to model the unpredictable and random fluctuations in concentration due to turbulent stirring and plume meandering. The empirical works by Eugene Yee et al. (Yee, Wang & Lien 2009; Yee 2009; Yee 2008; Yee & Biltoft 2004; Yee & Chan 1997) demonstrate that, in the absence of background, instantaneous concentration fluctuations fit very well the clipped-gamma probability density function (PDF). The clipped-gamma distribution is specified in terms of four parameters γ, *k**, *s* and λ as:

(Eq. 6.10)
$$f(\hat{c}) = \left(\frac{\hat{c}+\lambda}{s}\right)^{k*-1} \frac{\exp\left(-(\hat{c}+\lambda)/s\right)}{s\Gamma(k*)} + (1-\gamma)\delta(\hat{c})$$

where $\Gamma(\hat{c})$ is the gamma function, $\delta(\hat{c})$ is the Dirac delta function, and $\hat{c}$ is an array of all possible instantaneous concentrations *c*. The total PDF is composed of a mixed fluid part due to in-plume mixing of eddies containing the target substance (the first term on left-hand side), and an unmixed ambient fluid part (the second term on right-hand side) caused by plume meandering which produces intermittent periods of zero concentration for a fraction of time (1- γ). Although (Eq. 6.10) is specified in terms of four parameters, it can be uniquely modelled by the mean and the standard deviation (SD) of a series of readings including an additional constraint. This constraint (based on

previous studies (Yee et al. 1995; Yee et al. 1994; Yee et al. 1993)) provides a simple relation between the mean, the SD and the plume intermittency; thus, the intermittency factor γ is determined as:

(Eq. 6.11)
$$\gamma = \gamma(k^*, s, \lambda) = \frac{\Gamma(k^*; \lambda / s)}{\Gamma(k^*)} = \min\left(1, \frac{3}{K_{int}^{\ 2} + 1}\right)$$

where $\Gamma(v; c)$ corresponds to the incomplete gamma function and K_{int} is defined as:

(Eq. 6.12)
$$K_{int} = \frac{\sigma_{SD}}{M}$$

which is the ratio between the SD (σ_{SD}) and the mean (*M*) of the series of readings at a fixed position.

For a specified value of K_{int}, the parameters k^*, *s* and λ can be obtained solving a set of transcendental equations, thus making (Eq. 6.10) totally defined. The details for computing the parameters are given in (Yee, Wang & Lien 2009).

While the mean concentration decreases rapidly in the downwind direction (Eq. 6.9), the magnitude of the fluctuations decreases even more rapidly. As described in (Webster 2007), K_{int} (Eq. 6.12) is estimated to decrease as x^ϑ (where *x* is distance from the source and $\vartheta < 0$). Therefore, this parameter can be modelled as:

(Eq. 6.13)
$$K_{int} = K_0 \cdot \left(\frac{x}{x_0}\right)^\theta$$

In order to model the instantaneous concentrations due to a chemical plume at a certain distance from the source, the clipped-gamma distribution is used (Eq. 6.10). The mean value of the series of concentrations due to the plume is related to the GPM (Eq. 6.9) and the SD can be modelled using equations (Eq. 6.12) and (Eq. 6.13):

(Eq. 6.14)
$$\sigma_{SD} = K_0 \cdot \left(\frac{x}{x_0}\right)^\theta M$$

Since the PDF depends on the mean and the SD, and these parameters depend on the distance from the source, the PDF of the instantaneous readings contains information about the relative position between the sensor and the source.

Since concentrations fluctuations (intermittencies) decrease faster than the mean value with the downwind distance to the source, the plume becomes homogeneous faster than the mean concentration dilutes. It is important to note that this effect happens when no background is present and dispersion is dominated by turbulences (Jones & Thomson 2006; Yee, Wang & Lien 2009). In other studies, it has been reported that the mean concentration decreases in the vertical direction as SD increases, thus increasing intermittency (Crimaldi, Wiley & Koseff 2002). However, this last phenomenon is not considered in the present work, since only measurements at a constant height are obtained.

6.3.2 Background estimation

The problem of source localization in a complex environment under uncontrolled background signals has not yet been tackled in the published literature. In real scenarios such as forests, flat terrain, residential areas or urban environments, some kind of pollution or interfering substances are expected to be found. This problem becomes even more serious due to the common use of partially selective sensors like Metal Oxide Sensors, or Photo Ionization Detectors, which can provide large readings in the absence of the target substance merely due to the presence of interfering chemicals.

Although, as it has been reported in previous works (Hahn et al. 2009; Soriano et al. 2000; Martín, Cremades & Santabàrbara 1999), the meteorological conditions (wind conditions and atmospheric stability) could change within a timescale of several hours, or there might be changes in polluting emissions due to day-night cycles of human activity (including motor vehicles or factories), the background can be considered constant if the exploration time is comparatively short.

Based on the recorded sensor readings, our algorithm builds a statistical background model per cell. Assuming that the background changes slowly over time, and that no intermittencies are associated with it (all intermittency is considered to come from the plume), the SD of the series of concentrations is smaller than the mean value when no plume is present. Under these conditions, the clipped-Gamma distribution tends to a Gaussian distribution as it is shown in Figure 6.2. The main advantage of using a clipped-Gamma distribution rather than a Gaussian distribution for background modelling is that the former has only a positive range and thus has always a physical interpretation. Thus the background is modelled based on the clipped-Gamma distribution which can be estimated using two parameters.
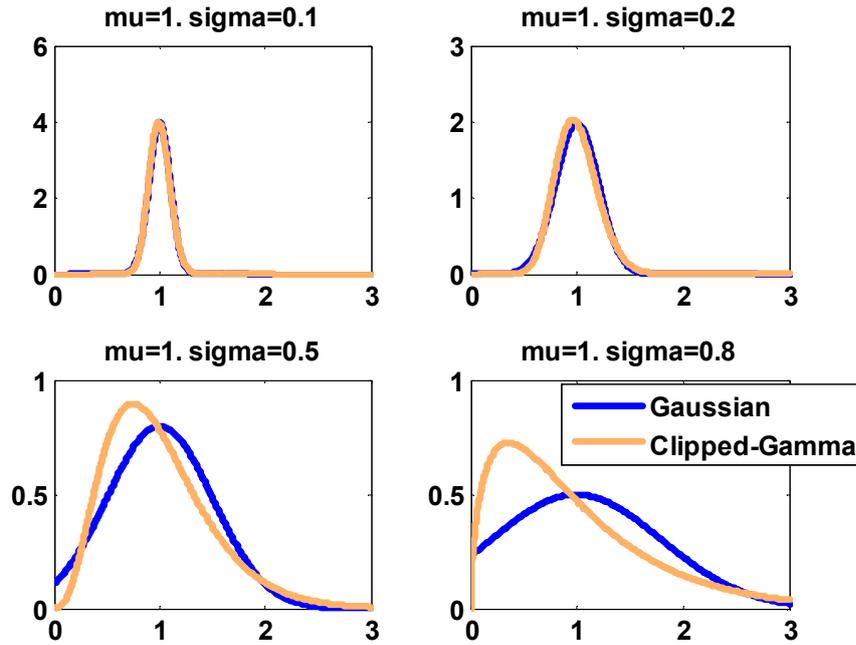
Figure 6.2. The clipped-Gamma distribution tends to a Gaussian distribution as the standard deviation decreases relative to the mean value.

As it has been described in previous work, in the absence of background, the dispersion of a chemical plume in a turbulent flow shows a highly intermittent nature with low or zero concentration for long periods of time separated by extremely high peaks of concentration (Webster 2007). However, in a real environment, in addition to this turbulent and chaotic dispersion of the plume, the background will appear overlapped with the plume (Eq. 6.8); therefore, given the instantaneous sensor readings, the true background needs to be extracted from this superposition of chemical signals.

As the robots explore the area, the instantaneous readings measured by each sensor are stored in separate buffers which contain the last N_b measurements. In the estimation of the two parameters of the background model, robust statistical estimators are needed in order to minimize the problem introduced by plume intermittency. The median and the median absolute deviance (MAD) are used rather than the mean and the SD, because the former are robust up to 50% of outliers (Gijbels & Hubert 2009). The background model for a specific cell is updated each time a sensor enters that cell. The median and the MAD are calculated using the measurements stored in that sensor's buffer. Then the statistical background model parameters of the cell are updated as follows:

(Eq. 6.15)
$$E = \frac{E_{old} + E_i}{2}$$

where E is the updated PDF parameter; E_{old} is the old PDF parameter; and E_i is the median or the MAD calculated from the measurements stored in the buffer. Under the assumption that the background evolves slowly over time, the updated values tend asymptotically to the true values as new measurements become available.

6.3.3 Source location probability map (concentration-based approach)

Following Pang's approach, the search area is divided into a grid with rectangular cells C_i (Figure 6.1). At the beginning of the exploration, the prior probability of source location is uniformly distributed among all cells with probability $1/N_c$. (In the presence of additional information, other prior distributions may be used to initialize the algorithm). Given that we measure a concentration at time t_k in cell C_j, we can calculate the source probability map based on this single reading. The two probabilities that the reading comes from the current background at cell C_j, or from a source at cell C_i, are weighted in order to calculate the final probability. From the simplest form of Bayes' theorem:

(Eq. 6.16)
$$P(A \mid c_j) = \frac{P(c_j \mid A){\cdot}P(A)}{P(c_j)}$$

using the law of total probability: $P(c_j) = \sum_{\varphi} P(c_j \mid A_\varphi){\cdot}P(A_\varphi)$, and considering the case of a binary partition (φ =1,2), where $A_1=A$ corresponds to the event "the concentration reading was caused by an emitting source upstream", $A_2=\overline{A}$ to the event "the concentration is caused by background levels" and c_j to the event "a certain concentration of chemical, c, has been measured in cell C_j", Bayes' formula can be rewritten as:

(Eq. 6.17)
$$P(A \mid c_j) = \frac{P(c_j \mid A){\cdot}P(A)}{P(c_j \mid A){\cdot}P(A) + P(c_j \mid \overline{A}){\cdot}P(\overline{A})}$$

Now the problem that we want to solve is: given a certain measured concentration c at cell C_j: what is the probability of having a source in cell C_i?
From (Eq. 6.17):

(Eq. 6.18)
$$P(A_i \mid c_j) = \frac{P(c_j \mid A_i) \cdot P(A_i)}{P(c_j \mid A_i) \cdot P(A_i) + P(c_j \mid \overline{A}_i) \cdot [1 - P(A_i)]}$$

where subindex *i* refers to cell C_i where the source may be located, and subindex *j* to the current cell C_j where the measurement was taken. c_j refers to a concentration reading in cell C_j.

We consider that a background of interfering substances is always present. From (Eq. 6.8): in the absence of plume, only the background component c_b is present; in the presence of plume, both components (c_b and c_p) are present and the concentrations are modelled with the convolution of the plume PDF and the background PDF. However, even in the presence of the plume, c_p may be zero due to plume intermittency. Plume PDF and background PDF are modelled using the clipped-Gamma distribution (Eq. 6.10). For the source, the PDF is computed from the mean and the SD of a hypothetical series of concentrations which would be observed in cell C_j due to a source at cell C_i. Since the readings are assessed independently and no time series is available, the mean is obtained from the GPM (Eq. 6.9) and the SD from (Eq. 6.14), as explained in section 6.3.1.

Taking these considerations into account, the previous probabilities have the following interpretation: $P(A_i)$ is the prior probability of the presence of a source at the cell C_i and is obtained from the PDF of the plume at cell C_j assuming that the source is at C_i; $P(c_j \mid A_i)$ is the probability that the measurement at cell C_j is due to addition of the background at cell C_j and a plume due to a source at cell C_i, and it is obtained by the convolution between the PDF of the plume and the PDF of the background at cell C_j. $P(c_j \mid \overline{A}_i)$ is the probability that the measurement of chemical at cell C_j is not due to a source emitting at cell C_i, therefore the measured concentration is due to the current background at cell C_j, and it is obtained from the PDF of the background at cell C_j. $S_{ij}(t_k) \equiv P(A_i \mid c_j)$ is the probability of having a source in cell C_i given that a certain amount of chemical was measured at cell C_j at time t_k.

In order to better illustrate how the algorithm works, we consider a possible scenario in the following example.
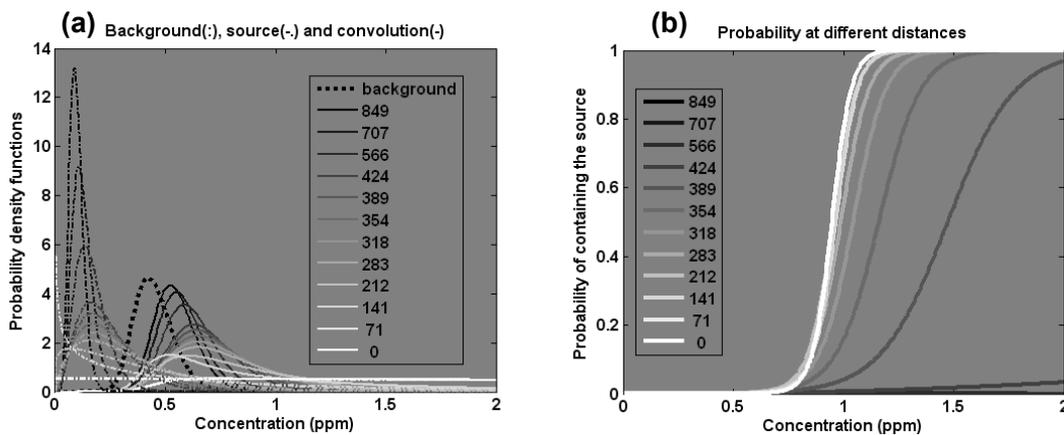
Figure 6.3. (a) PDF due to the background at current cell C_j (dotted line), PDFs due to the source (assumed to be at different distances) at current cell C_j (dash dotted line) and PDFs due to the convolution of background PDF and source PDFs. (b) Probability of there being a source at different upwind distances from the current cell C_j as a function of the measured concentration.

Figure 6.3 illustrates how the PDFs and the probability assigned to different cells look as a function of the measured concentration. It is supposed that a concentration has been measured in cell C_j having a certain background level (mean=0.45ppm and SD=0.09ppm in this example).

Figure 6.3(a) shows the background PDF at the measurement cell C_j, the plume PDFs at C_j assuming that the source is at different possible cells C_i (different distances from C_j), and the PDFs generated by the convolution of the plume PDFs and the background PDF. For each possible concentration, and for each cell C_i (each distance to C_j), a probability value can be extracted from the background PDF, the plume PDF and the PDF of the convolution. Please note that the background PDF is unique since we are considering one measurement, and the measurement cell C_j is fixed. Using these values in (Eq. 6.18) the probability S_{ij} (t_k) of having a source in C_i is obtained. This is illustrated in Figure 6.3(b), where S_{ij} (t_k) is depicted for different possible locations for the source (different distances to C_j) as a function of the measured concentration. It can be observed that, if the measured concentration is small or similar to the background level in C_j, the probability of there being a source in any cell C_i is close to zero. The probability starts to increase as the concentration becomes significantly higher than the background level. The slope of this increase depends on the distance from C_i to C_j. As the distance is increased, the slope decreases, consequently it is necessary to detect higher concentrations in order to come to believe that there is a source in C_i.

Although only some distances (cells C_i) are considered in Figure 6.3, equation (Eq. 6.18) can be calculated for each cell of the map ($i=1,...,N_c$), thus obtaining a probability for each cell. However, since the measured concentration in C_j could take any value, $\sum_{i=1}^{Nc} S_{ij}(t_k) = 1$ is not guaranteed. Moreover, it could be possible that all S_{ij} values ($i=1,...,N_c$) were below the initial value ($1/N_c$). Therefore, an offset is added to S_{ij} and the result is normalized to ensure a total probability of 1 when individual cell probabilities are added up. The nomenclature is henceforth the same as that used in Pang's work.

(Eq. 6.19)
$$\beta'_{ij}(t_k) = \frac{1}{N_c} + S_{ij}(t_k)$$

(Eq. 6.20)
$$\beta_{ij}(t_k) = \frac{\beta'_{ij}(t_k)}{\sum_{i=1}^{N_c} \beta'_{ij}(t_k)}$$

Now $\sum_{i=1}^{Nc} \beta_{ij}(t_k) = 1$ is guaranteed and $\beta_{ij}(t_k)$ calculated over all cells ($i=1,...,N_c$) gives the source probability map at time t_k based on a single measured concentration at cell C_j.

Using Bayesian theory (Papoulis 1984b) and following the same procedure explained in Pang's work (Pang & Farrell 2006), each new measurement can be incorporated recursively in order to update the source probability map. $\alpha'_{ij}(t_k) = P(A_i \mid B(t_k))$ is defined as the probability of cell C_i containing the source, given the sequence of concentrations $B(t_k)$ along the trajectory of the robots up to time t_k. Defining $P(A_i \mid D_j(t_k)) = \beta_{ij}(t_k)$, where $D_j(t_k)$ is the measured concentration at time t_k; $P(A_i \mid B(t_k))$ is computed from $B(t_{k-1})$ and $D_j(t_k)$, which are supposed to be independent events.

$$\alpha'_{ij}(t_k) =$$

$$P(A_i \mid B(t_{k-1}), D_j(t_k)) =$$

$$\frac{P(A_i B(t_{k-1}) D_j(t_k))}{P(B(t_{k-1}) D_j(t_k))} =$$

$$\frac{P(B(t_{k-1}) D_j(t_k) \mid A_i) P(A_i)}{P(B(t_{k-1})) P(D_j(t_k))} =$$

(Eq. 6.21)
$$\frac{P(B(t_{k-1}) \mid A_i) P(D_j(t_k) \mid A_i) P(A_i)}{P(B(t_{k-1})) P(D_j(t_k))} =$$

$$\frac{P(B(t_{k-1}) \mid A_i) P(A_i)}{P(A_i) P(B(t_{k-1}))} \frac{P(D_j(t_k) \mid A_i) P(A_i)}{P(D_j(t_k))} =$$

$$\frac{1}{P(A_i)} P(A_i \mid B(t_{k-1})) P(A_i \mid D_j(t_k)) =$$

$$\frac{1}{P(A_i)} \alpha'_{ij}(t_{k-1}) \beta_{ij}(t_k)$$

(Eq. 6.22)
$$\alpha'_{ij}(t_k) = N_c \cdot \alpha'_{ij}(t_{k-1}) \cdot \beta_{ij}(t_k)$$

where, if $\alpha'_{ij}(t_k)$ is computed over all cells ($i=1,...,N_c$), an updated source probability map is obtained recursively.

Furthermore, in order to minimize the effect of false alarms and to compare the binary-based and the concentration-based algorithms under exactly the same conditions, the source probability map is also filtered using the time constant τ in (Eq. 6.7): $\alpha_{ij}(t_k) = \dfrac{\alpha_{ij}(t_{k-1}) \cdot (\tau - 1) + \alpha'_{ij}(t_k)}{\tau}$.

The following table summarizes the notation used in this chapter.

C_i	Rectangular cell *i* within the search area (grid)
N_c	Number of cells in the grid area
m_x	Number of subdivisions of the grid area along the *x*-axis
m_y	Number of subdivisions of the grid area along the *y*-axis
L_x	Length of each cell along the *x*-axis
L_y	Length of each cell along the *y*-axis
c	Instantaneously measured concentration
c_b	Concentration contribution due to the background

c_p	Concentration contribution due to the chemical plume
$\bar{c}$	Mean concentration at a fixed location
q	Source strength or release rate
U_a	Mean wind speed in the downwind direction
σ_y	Diffusion coefficient in the crosswind direction
σ_z	Diffusion coefficient in the vertical direction
$\hat{c}$	Array of all possible instantaneous concentrations c
γ	Intermittency factor related to the chemical plume
M	Mean concentration of a series of readings at a fixed location
σ_{SD}	Standard deviation of a series of readings at a fixed location
N_b	Number of readings stored in the concentration buffer per each sensor
c_j	Measured concentration within cell j
A_i	Event "there is a chemical source within cell i"
$B(t_k)$	Sequence of measured concentrations along the trajectory of the robots until time t_k
$P(A_i)$	Prior probability of the presence of a chemical source within cell i
$P(c_j \mid A_i)$	Probability that the measurement within cell j is due to the addition of the background at cell j and a chemical plume due to a source within cell i
$P(c_j \mid \overline{A_i})$	Probability that the measurement of chemical at cell j is not due to a source emitting at cell i, thus c_j is due to the current background at cell j
$S_{ij}(t_k) \equiv P(A_i \mid c_j)$	Probability of having a source in cell i given that a certain amount of chemical was measured at cell j at time t_k
$\beta_{ij}(t_k)$	Normalized probability (over all cells) of having a chemical source within cell i based on a single measured concentration within cell j at time t_k
$\alpha'_{ij}(t_k)$	Normalized probability (over all cells) of having a chemical source within cell i based on the sequence of measured concentrations along the trajectory of the robots (index j) until time t_k
τ	Time constant for source probability map filtering
$\alpha_{ij}(t_k)$	Filtered $\alpha'_{ij}(t_k)$ probability using τ

Table 6.1. Summary of the notation used for the concentration-based approach.

6.4 Scenario description and simulation

6.4.1 Scenario description

In the present chapter, an adaptation of Pang's algorithm (Pang & Farrell 2006) (binary-based) and our implementation (which we name concentration-based and which uses continuous readings) are tested through realistic simulations.

The scenario envisioned considers atmospheric plumes in an urban area encompassing hundreds of thousands of square meters. The simulator devised allows a number of different functions: (i) generating a grid map of specified size where the actual localisation task takes place, (ii) changing the number of robots which will explore the area and their mean velocity, (iii) creating a certain background distribution in the area, (iv) simulating the emission of a chemical source with a certain release rate under (v) different wind and atmospheric conditions. It is considered that there is only one chemical source. As has been described in (Pang & Farrell 2006) and section 6.2, for computational feasibility, the search area is divided into rectangular cells C_j of a certain size (Figure 6.1). The sensors located in mobile robots can transmit their current position in the grid (e.g. using a GPS sensor) together with the chemical sensor readings, and they will explore the search area by moving across the cells performing random exploration. It is assumed that each robot mounts a single sensor. It is considered that the main task of the robots is not that of tracking the plume; rather the main interest is that of performing a mapping of the mean background level at the same time as updating a probability map for the source location using the available information. During the mission, the robots will sense the instantaneous chemical concentration at a certain sampling rate (a parameter which can also be set in the simulator); therefore a set of measurements will be available from each sensor in a separate buffer of N_b samples. The buffer size can also be set.

The model for chemical source dispersion is the time-averaged Gaussian plume model (GPM) (Turner 1994; Beychok 2005) which is the same model as that assumed by the source localization algorithm (section 6.3.1). Since the sensors measure instantaneous continuous readings and not time-averaged characteristics, concentration fluctuations need to be simulated. As has been described in section 6.3.1, concentration fluctuations due to the plume can be modelled using the clipped-gamma PDF and their statistical properties (mean, SD and intermittency) depend on the distance to the source. It is considered that plume dispersion is dominated by turbulence.

To generate realistic plume readings in each cell (different distances to the source), we have to generate random concentration signals –not only the PDF of the readings

needs to be taken into account, but also the expected power spectral density (PSD). Jones et al. in (Jones & Thomson 2006) present empirical studies of the PSD of chemical plumes obtained in short-range field dispersion experiments using an array of fast-response instruments deployed downwind from a point source of propylene gas. From the results presented in this work, we build a digital filter (FIR) to shape the PSD of the generated synthetic sequence of readings. The PSD is additionally band-limited with a maximum frequency given by half of the sensor sampling frequency.

Once we have defined the PDF and the PSD which will be used to model concentration fluctuations, we use the percentile transformation method (PTM) described in Papoulis (Papoulis 1984a) to generate a series of concentration fluctuations with the desired PDF and PSD.

Specifically, the procedure to generate realistic plume readings consists of the following steps: (i) generate a time series of Gaussian white noise, (ii) filter the previous time series with the designed FIR filter in order to achieve the desired PSD, (iii) and apply the PTM. This method is based on the following expression:

(Eq. 6.23)
$$c_i = F_c^{-1}\big(F_z(z_i)\big)$$

where z_i is a random sequence of Gaussian white noise having the desired PSD with cumulative distribution function (CDF) $F_z(z)$; c_i is the sequence of realistic readings in the cell with CDF $F_c(c)$. This CDF corresponds to the clipped-Gamma CDF which can be derived from the clipped-Gamma PDF (Eq. 6.10):

(Eq. 6.24)
$$F(c) \equiv \Pr\{C \le c\} = \int_{0^-}^{c} f(c')dc' = 1 - \frac{\Gamma\big(k*;(c+\lambda)/s\big)}{\Gamma(k*)}$$

and F_c^{-1} is the inverse clipped-Gamma CDF. The clipped-Gamma PDF (and its CDF) depends on the distance to the source. Its two parameters, the mean and the SD, are obtained from the GPM (Eq. 6.9) and from (Eq. 6.14), respectively; subsequently, these parameters are used to compute γ, $k*$, λ and s, as indicated in section 6.3.1 and explained in detail in (Yee, Wang & Lien 2009).

The background concentrations in each cell (with the desired mean and SD) are simulated using the same procedure and they are added to the series of plume readings. We consider that the background changes over time are relatively small compared to the exploration time.

This forms the final map of instantaneous concentrations which will be measured by the sensors. With each cell is associated its own series of instantaneous readings. An example of concentration fluctuations is shown in Figure 6.4. Simulated series of fluctuations at different fixed positions from the source are shown. When close to the source there are long periods of time where the sensor readings are equal to the background level. Additionally, the SD of the fluctuations decreases faster than the mean concentration, making it difficult to differentiate between plume and background far from the source. Figure 6.4 can be compared to Figure 1.22 which shows real concentration time series at two downwind distances.
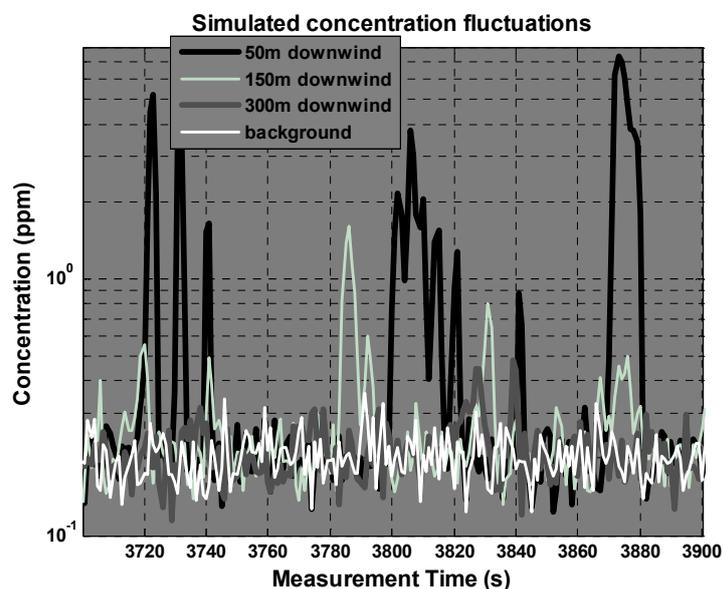
Figure 6.4. Simulated concentration fluctuations (under the specified conditions) at different fixed positions downwind from the source over a certain background level.

6.4.2 Evaluation of the algorithms by synthetic scenarios

In order to test both algorithms, a synthetic scenario is generated. A grid with a size of 1km x 1km is generated. The area is divided into cells of size 100m x 100m. A square sub-grid of lanes (100m separation between lanes) is interlaced over the main grid (Figure 6.5). The sensors will randomly explore the area over this last sub-grid of lanes which simulate streets within an urban environment (Manhattan style). Using this configuration, the movement of the sensors is constrained to certain points over the main grid, which is what we would expect in a real scenario with obstacles (buildings, parks, rivers...). A continuous Gaussian plume, emitting with a source strength q=2.90g/s, is placed in the grid with the source at the position (440m, 440m), which corresponds to coordinates (5, 5) on the rectangular grid. The dispersed substance is

170

acetone (molecular weight: 58.08g/mol) at one atmosphere pressure and 25ºC. This Gaussian plume distribution has been generated from (Eq. 6.9) as described in (Bakkum & Duijm 2005) with the plume being dispersed in a 2D plane at the same height as the sensors ($z=h=2m$). It is assumed that there is no deposition of the substance on surfaces. In the simulations, the wind field is constant with the wind speed at U_a=2.5m/s and the wind direction at 45º. The dispersion coefficients σ_y and σ_z depend on wind conditions and atmospheric stability which has been set to neutral ('D' on the Pasquill-Gifford scale (Bakkum & Duijm 2005)). Moreover, a mean background distribution is deployed over the area with a different mean level in each cell and with SD equal to 60% of the mean value in all cells (this 60% is based on our own recorded data using a PID sensor measuring outdoors in an open environment over a period of several hours); therefore all intermittencies come from the plume. The mean background level is different in each cell, but is constant over time. As described in section 6.4.1, series of concentration fluctuations are generated in each cell taking into account the wind field created, the atmospheric conditions and the background.

Five mobile sensors with a velocity of 15Km/h sense the area continuously. The sample time of the sensors is set to 3s, the detection limit to 0.1ppm and the sensor resolution to 0.01ppm. Thus, here we assume that the sensor time response is much faster than the typical 10 min time-average considered in the GPM (Eq. 6.9).
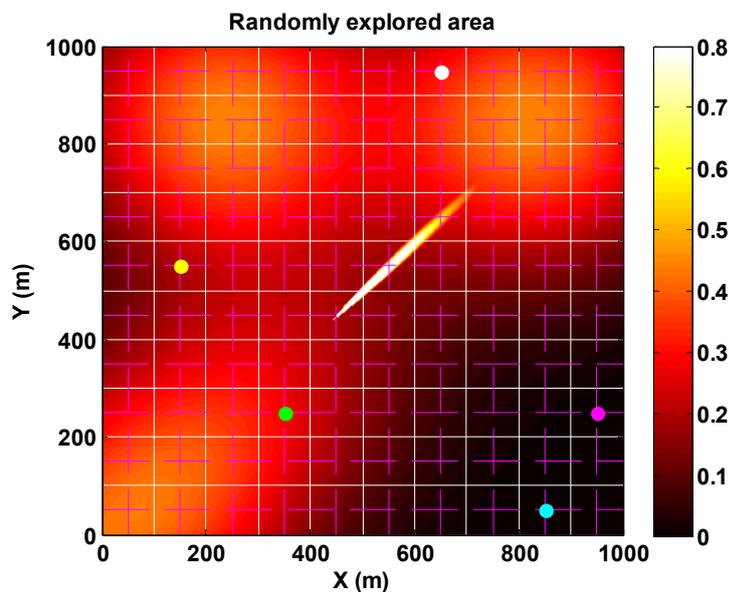
Figure 6.5. Randomly explored area (1km x 1km) using 5 sensors (coloured dots). Dashed lines (--) show the lanes over which sensors move; continuous lines (-) show cell boundaries. The source is located at (440,440) metres at cell (5, 5). Mean wind speed is set to 2.5m/s and wind direction to 45º.

The time constant τ related to probability maps filtering (binary-based and concentration-based) has been set to 450s. This time is related to the mean time taken by the group of robots to revisit a certain cell when performing random exploration. The simulation time was set to 300 minutes.

In the binary-based approach, it is assumed that the ratio of false alarms is very low, but the ratio of missed detection can potentially be very high, thus we define $\mu = 0.3$ (70% missed detections). However, the value of this parameter was not defined in the original work (Pang & Farrell 2006).

In the concentration-based approach, initially we assume that the source strength is known (2.90g/s), but since it is almost impossible to know this information in advance, we have studied the case when the assumed source strength is different from the real value within more than two orders of magnitude. For the simulation studies presented in this work, the parameters used in equations (Eq. 6.13) and (Eq. 6.14) are: K_0=2.5, x_0=50m and ϑ=-0.75, taking into account previous studies (Webster 2007). The buffer size (section 6.3.2) is set to N_b=50 which is neither too small (we have enough samples for computing the background parameters reliably) nor too high (samples from adjacent cells are taken and it is assumed that the background is similar).

Different simulations (three cases) were performed in order to compare the binary-based algorithm and our proposal, the concentration-based algorithm.

6.4.2.1 Case 1

The first simulation consisted in assessing the reliability of the binary-based approach when changing the concentration threshold, for two background levels (mean values: 0.05ppm and 0.45ppm). Thus, only in this case, the background was set constant over all cells. The reader should note that a background level is low or high depending on the source strength; therefore, studying the case where the background is low is equivalent to saying that the source is potent, and studying the case where the background is high is equivalent to saying that the source is weak. Ten random sensor trajectories were computed for each threshold level (30 values in the linear range between 0.05ppm and 3ppm) and for each trajectory a probability at the real source location is assigned by the algorithm after 300 minutes of exploration time. Finally, averaging over all trajectories, a mean probability is obtained as function of the concentration threshold.

6.4.2.2 Case 2

In the second simulation, both algorithms are compared for the same ten trajectories. Results show the probability evolution at real source location as the exploration time

increases. Moreover, the probability maps provided by both approaches after 300 minutes of exploration time can be compared and only in the case of the concentration-based algorithm is a mean background map recovered, which can be compared to the designed background distribution. Results are obtained for two background distributions (maximum mean values: 0.05ppm and 0.45ppm). Both distributions were the same except for a scale factor. The binary-based approach was tested using the optimum threshold level and the concentration-based approach using the same source strength as the one set in the simulator (2.90g/s).

6.4.2.3 Case 3

Finally, the third simulation shows the influence of the source strength in the overall performance of the concentration-based approach. Since it is difficult to know the source strength in advance, the performance of the algorithm has been assessed assuming different source strengths across more than two orders of magnitude in the range between 0.1g/s and 30g/s (the real value in the simulator being 2.90g/s). Results are shown for the same background distributions as in case 2 (section 6.4.2.2).

6.5 Simulation results and discussion

6.5.1 Case 1

The first result (Figure 6.6) shows the change in the overall performance of the binary-based approach as the concentration threshold is changed. Figure 6.6 shows the dependency for two different background levels: the former with a mean value of 0.05ppm (Figure 6.6(a)) and the latter with a mean value of 0.45ppm (Figure 6.6(b)). The mean probability assigned by the algorithm at the real source location is depicted as a function of the concentration threshold. It can be seen in the figure that there is a different optimal concentration threshold depending on the background level. Moreover, the optimum value is shifted to higher thresholds as the background level is increased. These optimal values have been found to be 0.15ppm and 1.48ppm, respectively. Relative to the background level, setting the threshold too low produces a high ratio of false alarms leading the algorithm to failure; on the other hand, setting the threshold too high could lead to an increase in false negatives with abnormal concentrations considered as non-detection events and a consequent worsening of the overall performance. This can be critical if the source which is to be detected is weak – this being equivalent to the case with a high background level. In Figure 6.6(b), it is observed that setting the threshold either too low or too high causes the algorithm to

fail, since the probability assigned to the real source location is below the equiprobable value of $1/N_c$ assigned initially to every cell.
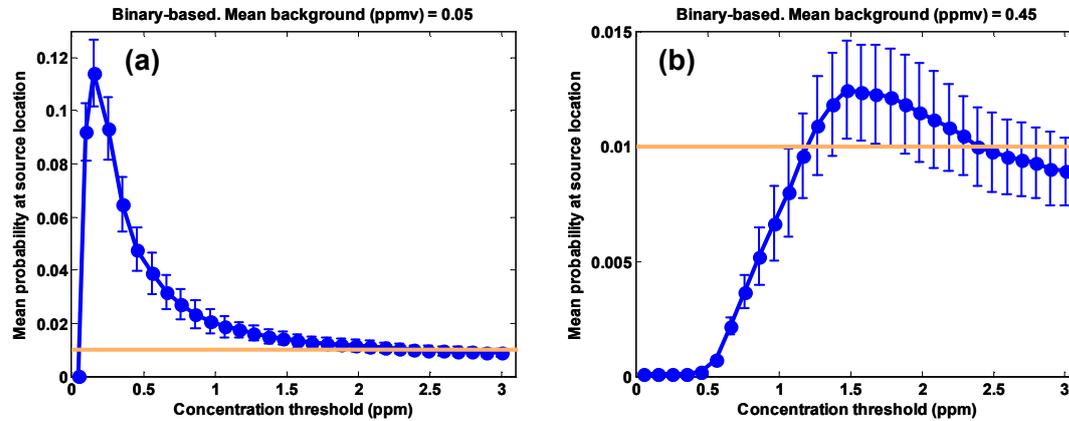
Figure 6.6. Dependency of the assigned mean probability (averaged over 10 different random trajectories) with the concentration threshold at source location by the binary-based approach. (a) Low background level (mean = 0.05ppm and SD = 0.03ppm). (b) High background level (mean = 0.45ppm and SD = 0.27ppm). The orange line shows the initial equiprobable value ($1/N_c$) assigned to every cell.

The main problem using the binary-based approach is that the threshold needs to be set arbitrarily if no information about the background is available, and this background can be different in various areas within the exploration zone. A priori, we do not know whether the threshold is too high or too low, but even if we knew this, the background could evolve over space and time and the threshold would need to be adjusted continuously. The concentration-based algorithm removes the necessity of any threshold because, instead of adjusting the threshold level, our approach builds a background model for each cell. This background model allows us to distinguish between the background and the plume without using any threshold and is updated recursively.

6.5.2 Case 2

Results from the second simulation are shown from Figure 6.7 to Figure 6.14. Figure 6.7 to Figure 6.10 show the results obtained with a background distribution having a maximum mean value of 0.05ppm. Figure 6.11 to Figure 6.14 show the results obtained with a background distribution having a maximum mean value of 0.45ppm. With a low background level (0.05ppm), Figure 6.7 shows the probability assigned by both algorithms at the real source location as exploration time increases. Figure 6.7(a)

shows the increase in the probability for ten different random trajectories. Figure 6.7(c) (concentration-based) and Figure 6.7(d) (binary-based) show the increase the mean probability (averaged over all trajectories) with the margins related to two standard deviations. Figure 6.7(b) shows the results of a statistical test (the Wilcoxon test). The test is applied at each time, considering the probabilities at each time of each algorithm as two variables (10 points for each variable from the trajectories). A p-value is returned by the test. The closer the value is to 0, the more doubt there is of the validity of the null hypothesis (that the medians are equal). It can be seen how the concentration-based approach performs significantly better and how this increases with the exploration time, even though the binary-based algorithm has been optimized with the best concentration threshold (0.15ppm).
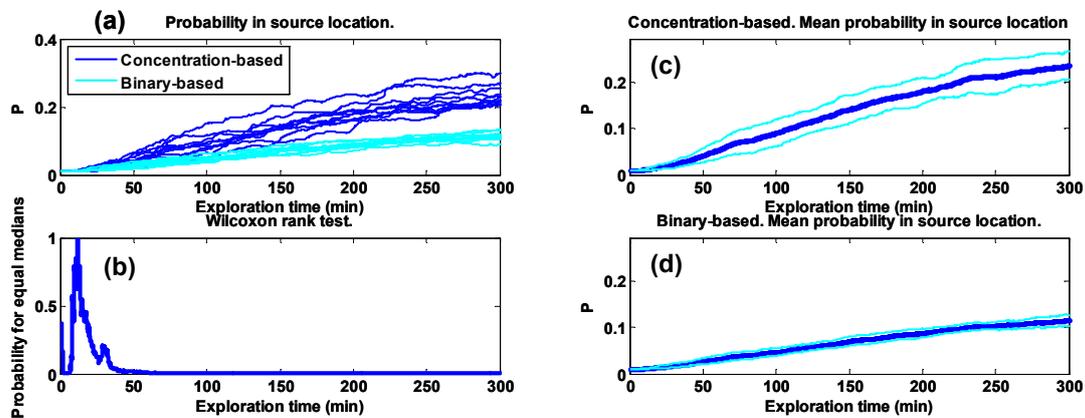
Figure 6.7. Probability at real source location with a maximum mean background level of 0.05ppm, as exploration time is increased. (a) Probability for 10 random trajectories using both approaches. (b) Wilcoxon test applied to both approaches. (c) Mean probability (averaged over all trajectories) and confidence intervals within two standard deviations for the concentration-based approach. (d) Mean probability (averaged over all trajectories) and confidence intervals within two standard deviations for the binary-based approach.

Figure 6.8 shows the comparison between the mean probability maps (averaged over all trajectories) provided by both algorithms after 300 minutes of random exploration. The probability assigned to the real source location is higher using the concentration-based approach (P=0.24 in Figure 6.8(b)) as compared to the binary-based approach (P=0.11 in Figure 6.8(a)). Moreover, the probability is spread among a lower number of cells in the wind direction.
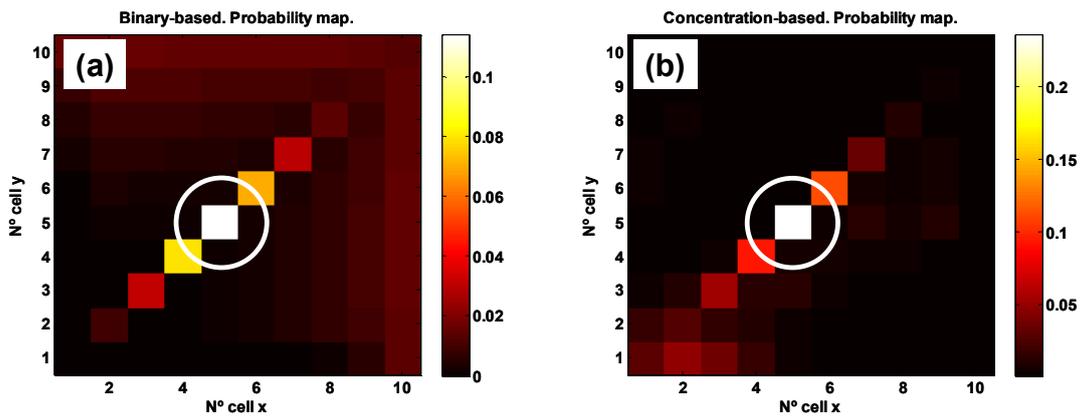
Figure 6.8. Mean probability maps (averaged over all trajectories) after 300 minutes of random exploration with a maximum mean background level of 0.05ppm. Source location at (5,5). (a) Binary-based approach. (b) Concentration-based approach.

Figure 6.9 shows a comparison between the real designed background distribution and the recovered background distribution by the concentration-based approach after 300 minutes of random exploration. The estimated background distribution is quite similar to the original one. This estimation is only possible using the concentration-based algorithm.
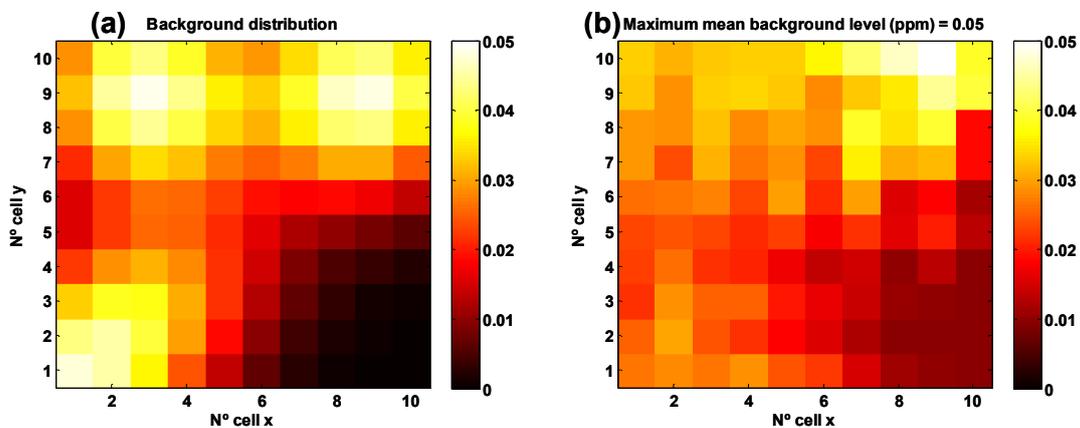
Figure 6.9. (a) Designed background distribution with a maximum mean background level of 0.05ppm. (b) Recovered mean background distribution (averaged over all trajectories) using the concentration-based algorithm.

Figure 6.10 shows the root mean squared error (RMSE) between the recovered background and the designed one. It can be seen that after 50 minutes of exploration the error remains approximately constant. The RMSE is computed as:

(Eq. 6.25)
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_C}\left(B_i - \hat{B}_i\right)^2}{N_C}}$$

where B_i is the designed mean background level at cell C_i; $\hat{B}_i$ is the recovered mean background level; and N_c is the number of cells in the map.

Mean differences between true and recovered background
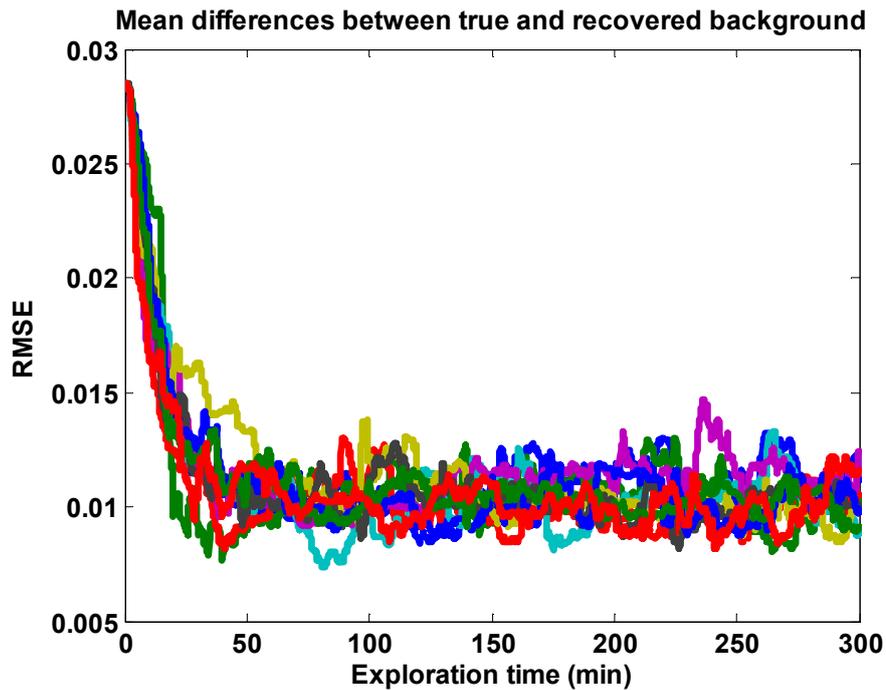
Figure 6.10. Root mean squared error (RMSE) for the differences between the designed background distribution (maximum mean level of 0.05ppm) and the recovered one, as the exploration time increases. The graph shows the RMSE evolution for the 10 trajectories.

With a high background level (0.45ppm), which is the same as saying that the source strength is small compared to the background, Figure 6.11 shows similar results to that shown in Figure 6.7. However, now the differences between the performances of both approaches are higher. Although the probability at real source location is now increasing more slowly (compared to Figure 6.7) due to a high background level, the slope is higher in the concentration-based approach as compared to the binary-based algorithm. The binary-based approach performs well under the assumption that no false alarms arise. This is shown in Figure 6.7 where, due to a low background level, the number of missed detections and false alarms arising from the background are small, and the binary-based algorithm performs a little worse compared to the concentration-based approach if the optimum threshold can be identified. Nevertheless, such an assumption is far from the truth in a real scenario where

pollution and interfering substances are expected to be found. Additionally, this can also be the case when the source to be detected is weak. This is shown in Figure 6.11 where, due to a high background level (or weak source), the number of false alarms is higher, thus forcing setting the threshold higher, which leads in turn to an increase in the number of missed detections and worsens dramatically the performance of the binary-based approach. The performance of the concentration-based approach has also been worsened due to the low background levels, however is significantly more robust than the binary-based algorithm, although the latter has been optimized with its optimum concentration threshold.
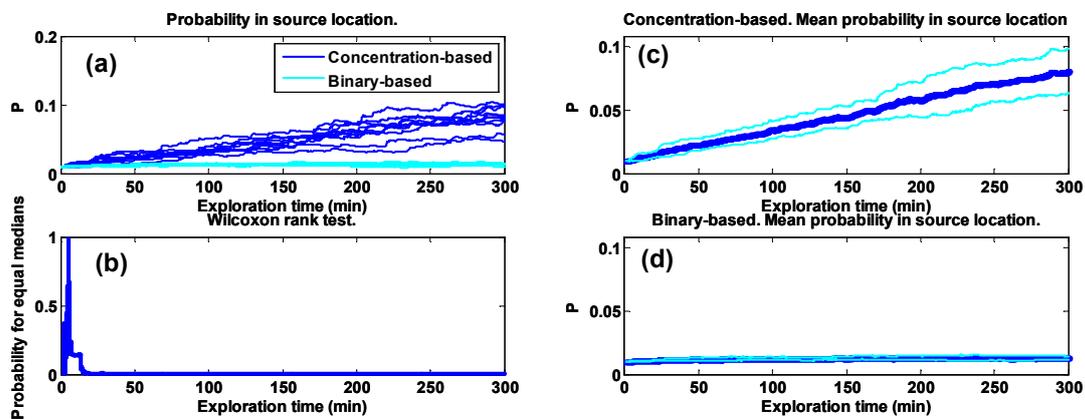
Figure 6.11. Probability at the real source location with a maximum mean background level of 0.45ppm, as exploration time is increased. (a) Probability for 10 random trajectories using both approaches. (b) Wilcoxon test applied to both approaches. (c) Mean probability (averaged over all trajectories) and confidence intervals within two standard deviations for the concentration-based approach. (d) Mean probability (averaged over all trajectories) and confidence intervals within two standard deviations for the binary-based approach.

Figure 6.12 shows the probability maps obtained with a high background level. It is seen that the threshold is very high, which minimizes the number of false alarms but increases the number of missed detections, thus the binary-based estimation is very uncertain at the source location (position (5, 5), Figure 6.12(a)). Therefore, the probability is spread over the cells in the grid, decreasing the probability at real source location (P=0.013). Figure 6.12(b) shows the robustness of our algorithm which minimizes false alarms and missed detections. The concentration-based algorithm tends to increase the probability at the real source location while the robots are performing random exploration. After 300 minutes of random exploration, the probability assigned to real source location was 0.08. In this approach, false alarms arising from the background can correctly be assigned lower weights in the probability

calculations because the algorithm has created a background model per each cell and a dispersion model for the plume. Additionally, these models allow minimizing the number of missed detections.
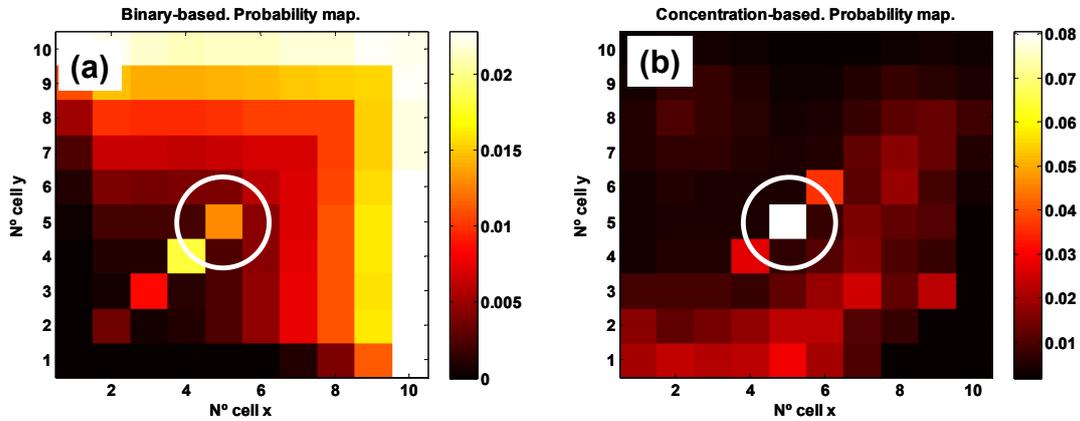
Figure 6.12. Mean probability maps (averaged over all trajectories) after 300 minutes of random exploration with a maximum mean background level of 0.45ppm. Source location at (5,5). (a) Binary-based approach. (b) Concentration-based approach.

Figure 6.13 shows that the concentration-based approach is also able to recover properly the background distribution in this case (maximum mean background level: 0.45ppm).
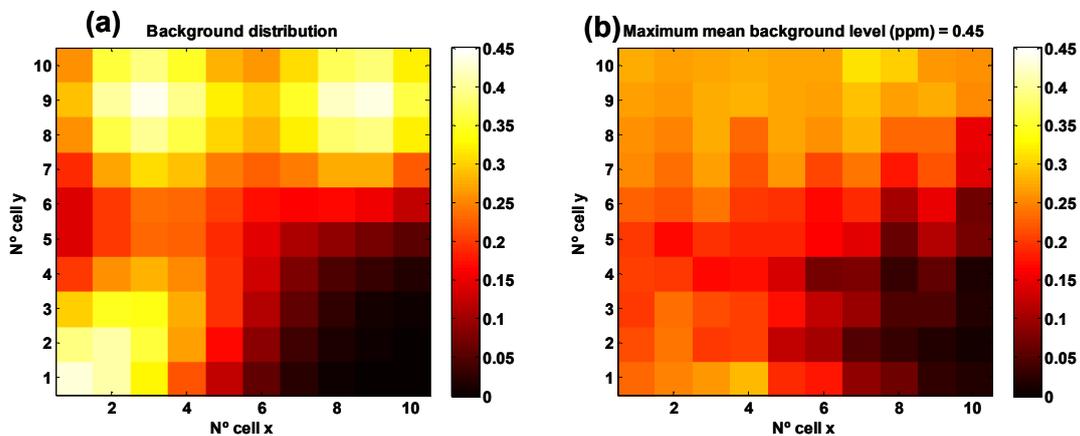
Figure 6.13. (a) Designed background distribution with a maximum mean background level of 0.45ppm. (b) Recovered mean background distribution (averaged over all trajectories) using the concentration-based algorithm.

Figure 6.14 shows how the RMSE (Eq. 6.25) evolves as the exploration time increases and that the error remains approximately constant after 50 minutes. The error is higher compared to that of Figure 6.10.
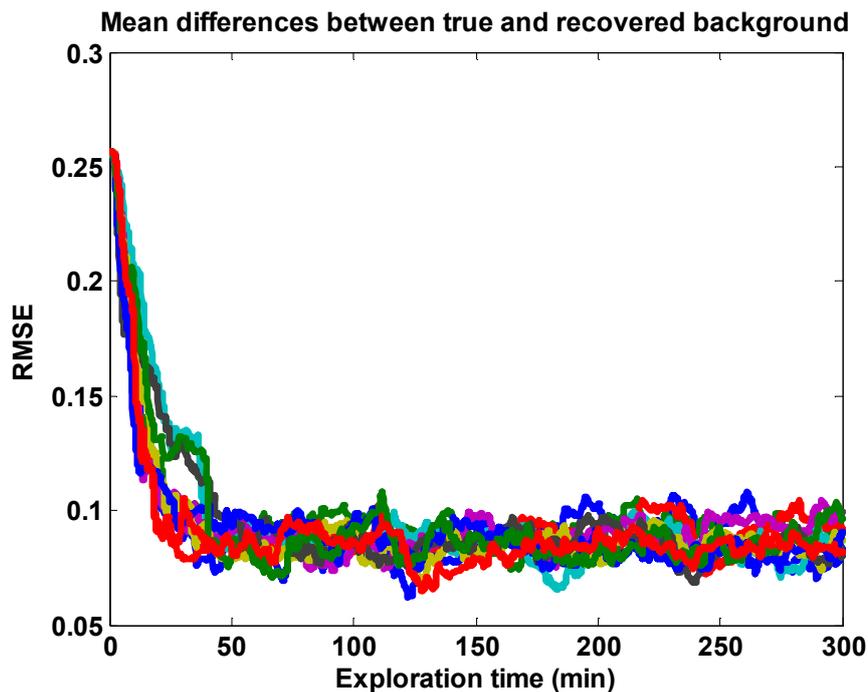
Figure 6.14. Root mean squared error (RMSE) for the differences between the designed background distribution (maximum mean level of 0.45ppm) and the recovered one, as the exploration time increases. The graph shows the RMSE evolution for the 10 trajectories.

6.5.3 Case 3

The results obtained from the third simulation are displayed in Figure 6.15. It shows the mean probability at the real source location as a function of the source strength assumed by the concentration-based approach. Results show the decrease in the overall performance as the assumed source strength deviates from the real value (2.90g/s). It is observed that the selection of the source strength becomes more critical when the background level is low (or the source is potent compared to this background), but, if the source strength could be estimated somehow, the concentration-based approach performs much better. When the source to be detected is weak (or the background is comparatively high, Figure 6.15(b)), the selection of the source strength is not so critical in the range studied. It can be observed that the concentration-based approach is more robust against false alarms and missed detections (even though the optimum source strength is not selected) as compared to the binary-based approach which performs badly even setting its optimum threshold.
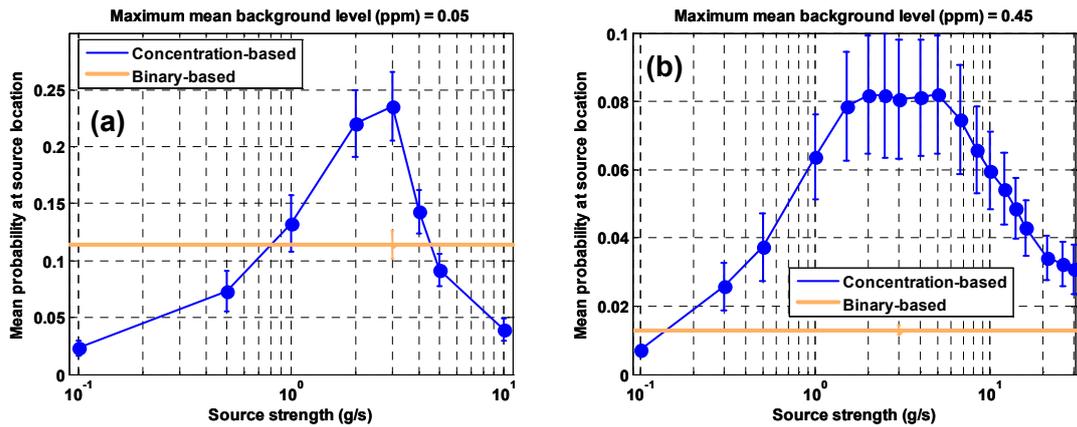
Figure 6.15. Mean probability (averaged over 10 trajectories) at real source location after 300 minutes of random exploration as a function of the source strength assumed by the concentration-based approach. Error bars show confidence levels within two standard deviations. (a) Results with a maximum mean background level of 0.05ppm. (b) Results with a maximum mean background level of 0.45ppm.

It is important to say that, although the binary-based algorithm works without assuming explicitly any source strength, setting an optimum threshold is only possible when collecting real measurements which implicitly contain information about the source strength and the background. However, this optimum threshold might be different depending on the explored cell. In the case of the concentration-based approach, the background is estimated by the algorithm and, if this background is low, the source strength needs to be known to within one order of magnitude. If the background is high, the algorithm behaves more robustly in the range studied as compared to the binary case except in the first case (the assumed source strength is very small).

We consider that estimating the source strength provides a more promising direction for future research rather than trying to optimize the threshold in the binary-based approach, especially in cases where a weak source has to be detected. In the concentration-based approach no information is dismissed since all readings are used, which allows the minimization of the number of false alarms and missed detections.

6.6 Real-world data using a mobile robot

In order to validate and support the obtained results from simulations, real-world data is required. In this section, the materials and the experimental set-up built to obtain these data are described.

The experiments were carried out in the *Mobile Robotics and Olfaction Lab* (Örebro, Sweden).

6.6.1 Materials

A mobile robot (Figure 6.16) was used in all the experiments carried out. It incorporates the following elements:

- *Embedded* **PC** with Ubuntu (Linux) as operating system.
- *Router* **WIFI** provides wireless communication with the robot.
- *Laser scanner* SICK LMS 200. Robot localization indoors.
- *Xsens MTI-G* with GPS antenna. Robot localitzation outdoors.
- **Ultrasonic Anemometre**. Wind measurements (speed and direction) at sampling frequency 2Hz in the range between 0 and 60m/s with resolution 0.01m/s.
- **Gas sensor array**. Sensor array (model Figaro TGS) which includes 4 MOXs (*Metal Oxide Sensor*) and one electrochemical sensor for CO_2 detection.
- **Photoionization detector** (PID) ppbRAE 3000 (RAE systems) with 10.6eV UV lamp. It provides calibrated measurements (chemical concentrations in ppmv) at sampling frequency 1Hz. It incorporates an internal pump at 500ml/min.
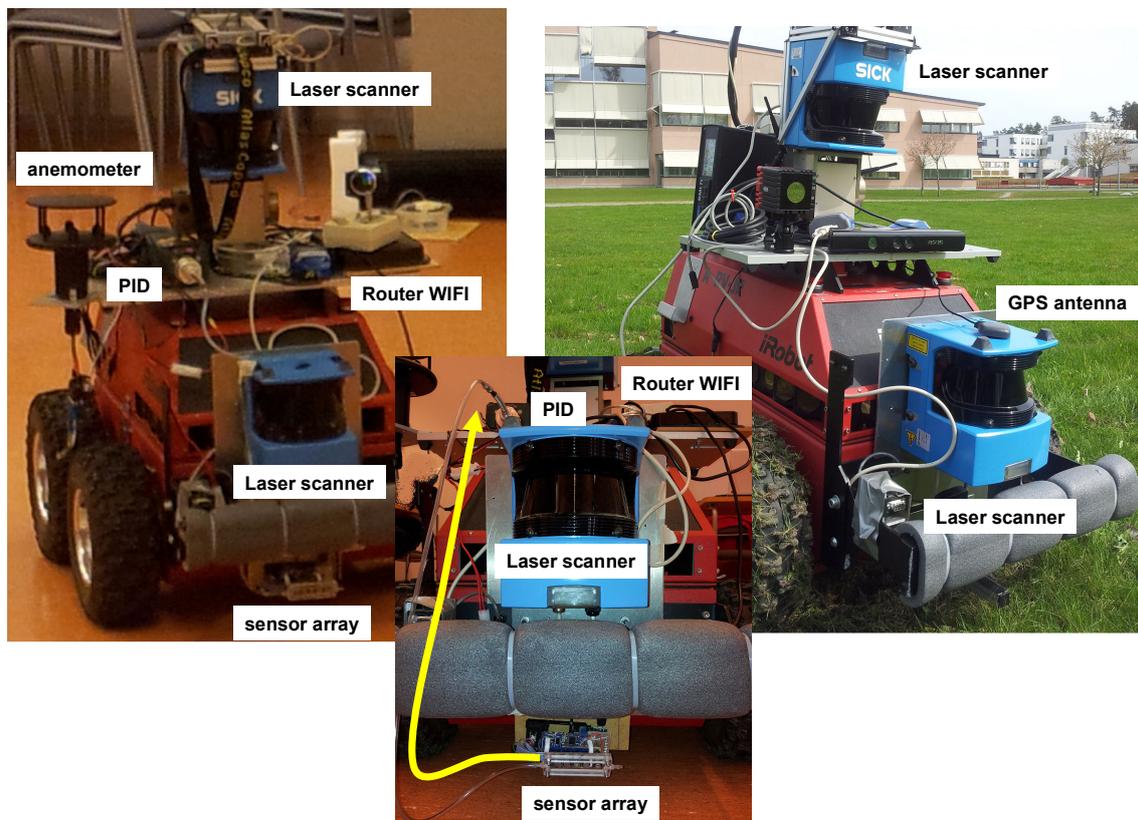
Figure 6.16. Mobile robot and instrumentation used for the chemical source localization experiments.

6.6.2 Evaluation of the algorithms by real scenarios

6.6.2.1 Experimental set-up

Twenty experiments were carried out under forced ventilation, indoors (Figure 6.17) and outdoors (Figure 6.18), with and without obstacles and with one or two chemical sources at the same time.

The selection of chemical substances was based on commercial availability, hazard, toxicity, chemical volatility and sensitivity; thus acetone, ethanol and 2-propanol were chosen. The substances were placed in a container in liquid state. A small pump connected to a plastic tube was used to generate bubbles in the liquid and thus to favour the vaporization of the substances (Figure 6.17). Two fans are placed in one of the boundaries of the exploration area in order to create a forced airflow (turbulent). It is assumed that the atmospheric stability is neutral.

The robot was programmed to explore the area randomly (velocity $\approx$ 0.2m/s) stopping for 30 seconds at different fixed positions. The robot captures wind information (speed and direction) and chemical readings continuously. The area size was $7x6m^2$ (indoors) and $9x7m^2$ (outdoors) and the cell size was set to $1x1m^2$ in order to have enough measurements to compute and update the background model of each cell.

It is important to note that, although measurements from the gas sensor array were available, our source localization algorithm requires chemical concentrations (calibrated measurements) in order to update the source probability map. Since specific data processing would be needed to obtain calibrated concentrations from the sensor array and the objective is to validate the simulation results, for simplicity, only the PID measurements are taken into account. This sensor is calibrated using gas isobutylene and a correction factor must be applied to obtain the concentration of different target gases. A table with correction factors is available from the manufacturer (RAE systems).
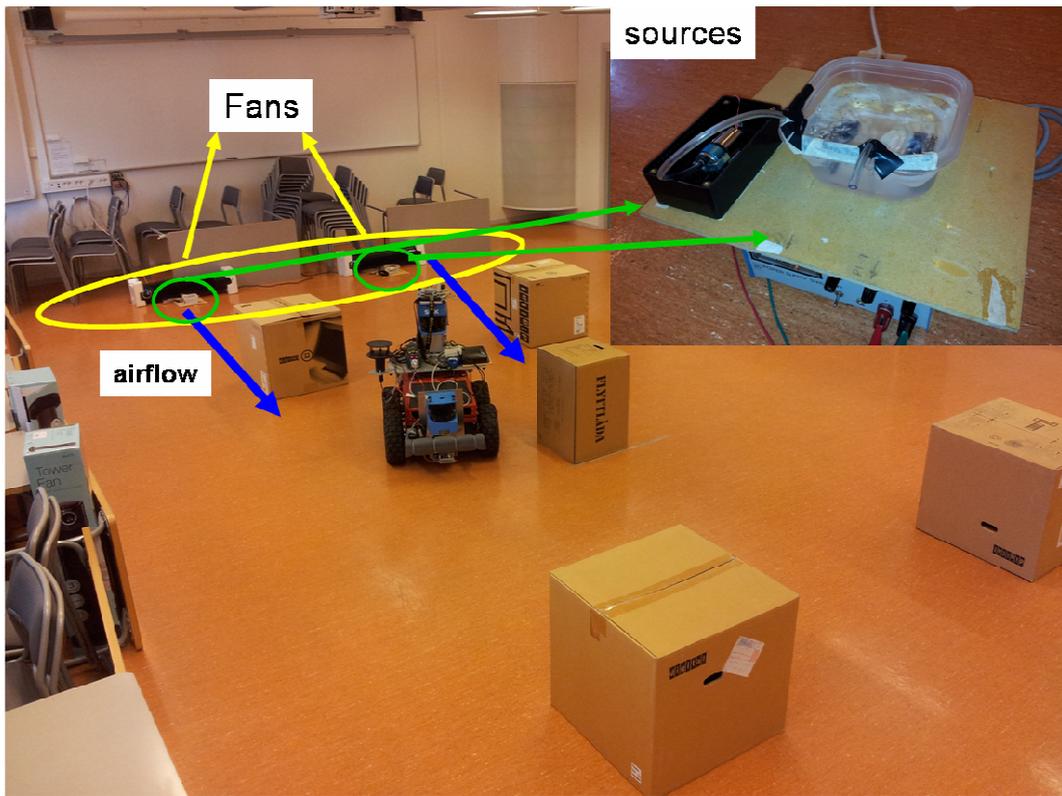
Figure 6.17. One of the experiments performed indoors with two chemical sources and plume dispersion in the presence of obstacles.
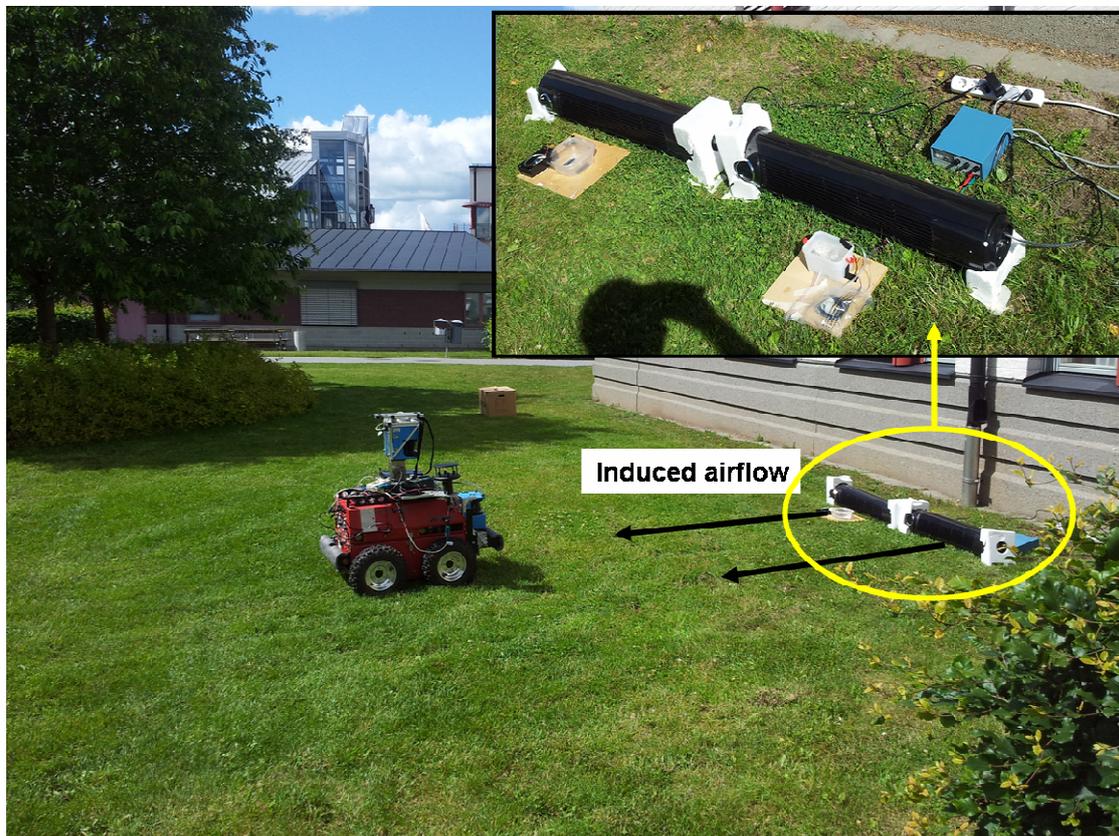
Figure 6.18. One of the experiments performed outdoors.

6.6.2.2   Algorithms evaluation

Since the objective is to test the performance of both approaches and multiple experiments were carried out, we should consider some figures of merit in order to analyze and summarize the results.

In the experiments with one chemical source, three figures of merit are considered:

- Final probability at the real source location.
- Distance between the probability maximum and the real source location.
- Overall entropy of the final probability map, defined as:

(Eq. 6.26)
$$H = -\sum_{i=1}^{N_{cells}} P_i \cdot \log_{Ncells}(P_i)$$

P_i is the probability assigned to cell *i* in the grid map, and N_{cells} is the total number of cells. Taking N_{cells} as the basis of the logarithm ensures that ($1 \geq H \geq 0$) and different experiments can be compared. Lower values of *H* correspond to a higher certainty regarding the source location.

A Wilcoxon test at the 0.05 significance level is performed in the different figures of merit so as to evaluate if the concentration-based approach performs significantly better than the binary-based approach.

Due to the fact that both the binary-based and the concentration-based algorithms assume that there is only one source in the search area, the experiments containing two sources are analyzed separately. In these experiments, therefore, the distance to the real source location and the probability at the real source location cannot be taken as figures of merit.

As it is shown in section 6.5, the performance of the algorithms mainly depends on the selection of their parameter: the concentration threshold (binary-based) and the source strength (concentration-based). For each experiment, each parameter is optimized independently and it is selected from a parametric sweep after testing 20 different values. It is important to remark that, in this section, we are interested in testing the performance of both approaches under the optimum conditions rather than finding a method to identify the optimum parameters. Finding such a method is an open issue for further work.

6.7 Real-world data results and discussion

6.7.1 Single-source experiments

The results of a representative experiment are shown in this section (indoors, with obstacles, 2-propanol).

The instantaneous measurements from the PID and the anemometer at each point in the random trajectory of the robot are shown in Figure 6.19. In both approaches, (binary-based and concentration-based) a probability map of the source location is updated as new measurements are available. Figure 6.20 shows the probability maps at the end of the exploration, obtained with the optimum parameters.
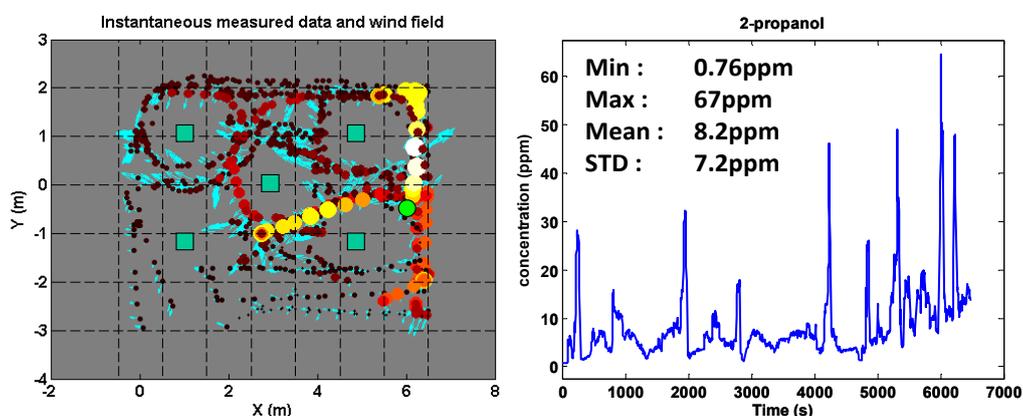
Figure 6.19. Left: Instantaneous concentrations measured with a photo-ionization detector (PID), shown in a specific heatmap (higher concentrations are additionally indicated by larger dots) together with measured mean wind vectors (cyan). The source is located at (6, -0.5)m, depicted using a green dot. Obstacles appear as green squares.
Right: Instantaneous concentration measurements from the PID during the full experiment.
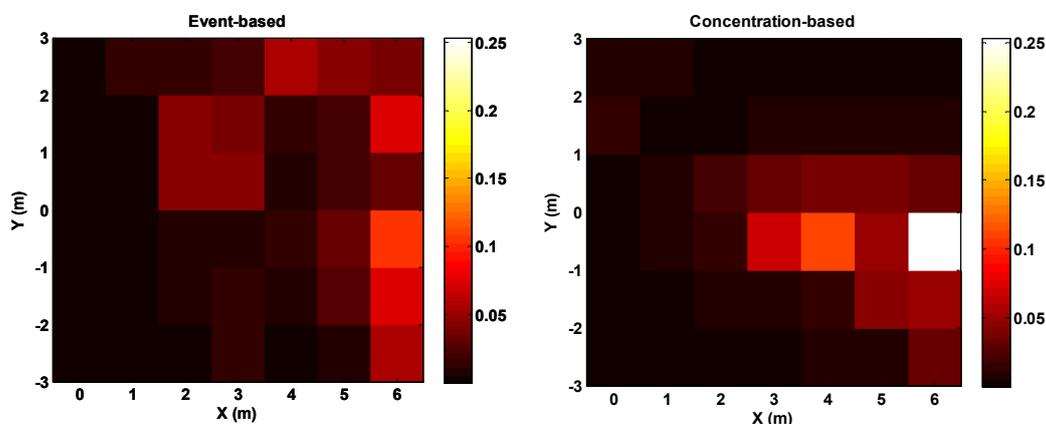
Figure 6.20. Final probability maps provided by the binary-based (left) and concentration-based (right) approach using optimized parameters. The source is located at (6, -0.5) m.

From these results, the three figures of merit can be computed: the probability at the real source location (P_{bin} and P_{conc}), the distance between probability maximum and source location (D_{bin} and D_{conc}), and the entropy (Eq. 6.26) over the final probability map (H_{bin} and H_{conc}).

This is done for the 12 experiments with a single source, obtaining the following boxplots:
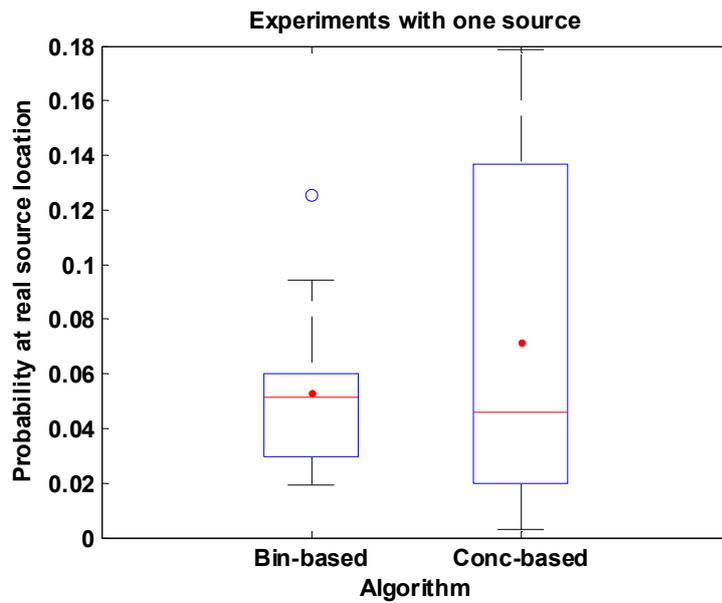
Figure 6.21. Final probability at real source location. Boxplot of the distribution of the results for the binary-based and concentration-based algorithms. The top and bottom of the box are the 25th and 75th percentiles. The point within the box is the mean and the horizontal line is the median. The whiskers extend to the most extreme data points which are not considered outliers, which are represented as circles ('o').
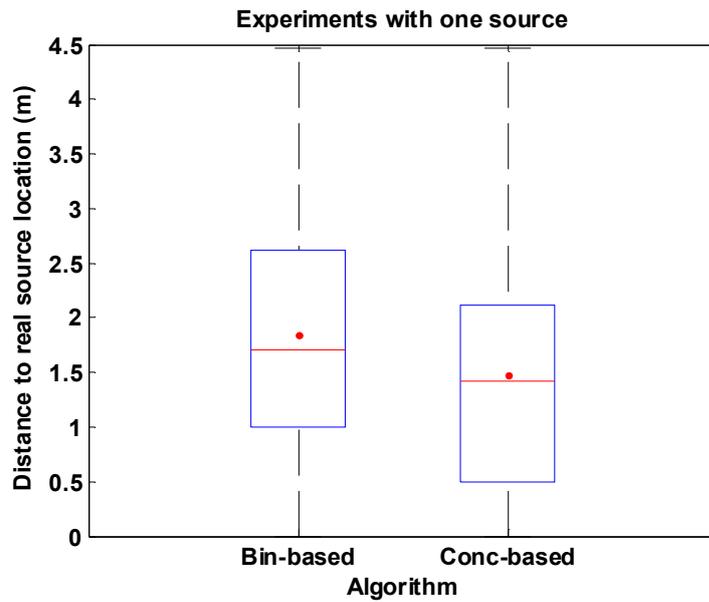
Figure 6.22. Distance to real source location. Boxplot of the distribution of the results for the binary-based and concentration-based algorithms. The top and bottom of the box are the 25th and 75th percentiles. The point within the box is the mean and the horizontal line is the median. The whiskers extend to the most extreme data points which are not considered outliers.
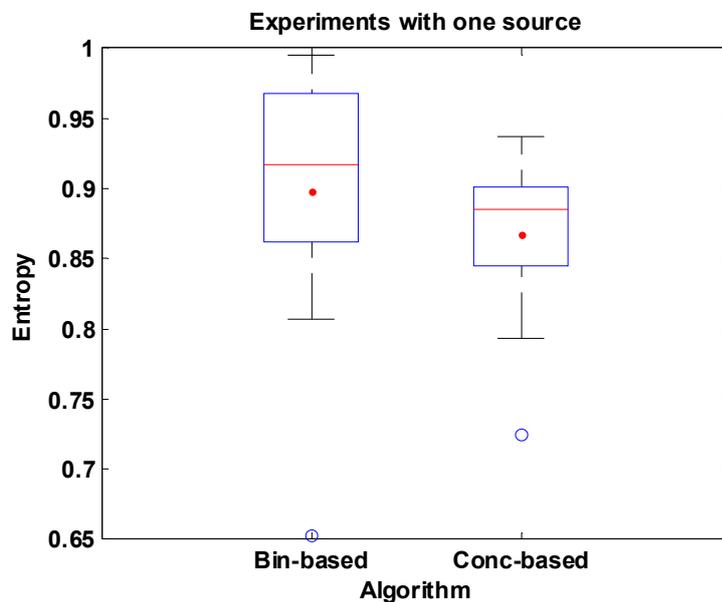
Figure 6.23. Entropy. Boxplot of the distribution of the results for the binary-based and concentration-based algorithms. The top and bottom of the box are the 25th and 75th percentiles. The point within the box is the mean and the horizontal line is the median. The whiskers extend to the most extreme data points which are not considered outliers, which are represented as circles ('o').

Although these results confirm that the concentration-based approach behaves more robust on average, the results from the Wilcoxon test suggest that its performance is not significantly better (p-values: 0.08, 0.5, 0.98, for entropy, distance and probability at real source location, respectively).

As it is extracted from the results from simulations, the concentration-based approach provides similar results compared to those from the binary-based approach when the background of interfering signals is low (which is the same as saying that the source is potent) (Figure 6.7). On the other hand, when the source to be detected is weak (high background level) the concentration-based algorithm provides much better results (Figure 6.11). In the experiments with the mobile robot which were carried out there is no real background of interfering chemicals or it is very low, thus the algorithms are tested in favourable conditions for Pang's approach. Therefore, the obtained results using real-world data are in agreement with the simulation results (Figure 6.7).

Nowadays, a database of real-world data designed to validate and test the case with a high background level (or weak source) has not been created. Under these conditions it would be expected that the concentration-based approach would perform much better as shown in the simulation results (Figure 6.11).

6.7.2 Two-source experiments

One could think that a second source could be considered as a high chemical interference in the background and then test the algorithms under these conditions. However, a second source cannot be considered as interfering because statistically would be characterized as a source by the concentration-based algorithm. This approach relies on statistics to discriminate between plume and background. Since the typical sensors are not selective (PID included), they react to a high number of substances, and moreover, since the algorithm assumes that there is only one source, assuming statistical distributions is necessary for the discrimination.

It is assumed that a chemical source produces peaks of high concentration combined with long periods of time where no chemical is detected. Moreover, this depends on the distance to the source and there is a statistical model which takes it into account. The algorithm builds the background model from the instantaneous readings from the sensor (plume + background) and relies on robust statistics (median and MAD) to extract the background from the readings. If there is a second source, the algorithm will "see" it as a source, since statistically it is characterized as it, and the background will be what is "behind", supposed to be something smooth with no abnormal peaks of high concentration; in fact these peaks are discarded by the median and the MAD since they are considered to come from the chemical source.

The results of a representative experiment using two sources are shown in this section (outdoors, no obstacles, acetone and ethanol).
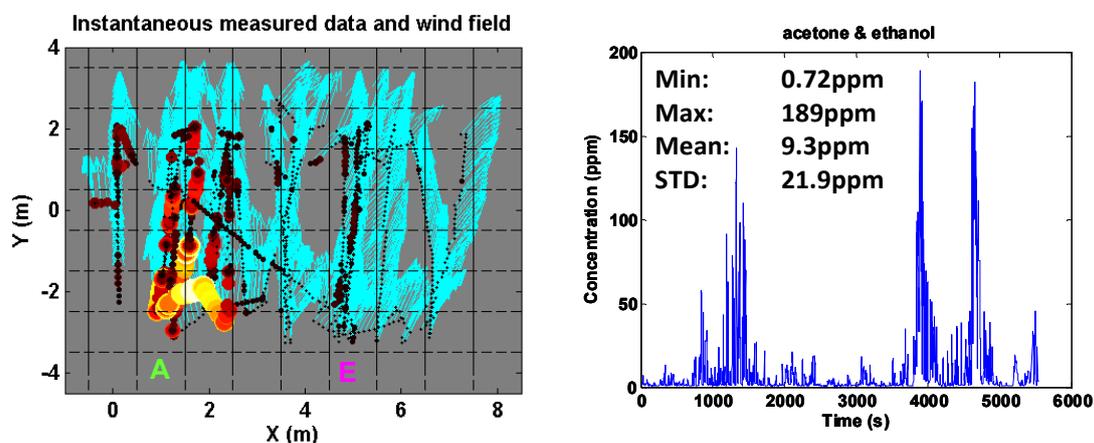
Figure 6.24. Left: Instantaneous concentrations measured with a photo-ionization detector (PID), shown in a specific heatmap (higher concentrations are additionally indicated by larger dots) together with measured mean wind vectors (cyan). The sources are located at (1, -4)m for acetone ('A') and (5, -4) for ethanol ('E'), depicted using coloured letters.
Right: Instantaneous concentration measurements from the PID during the full experiment.
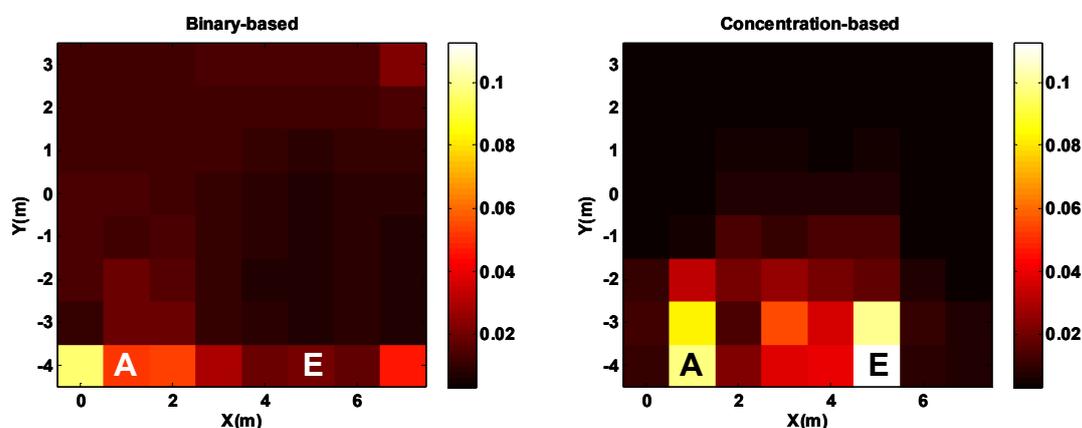
Figure 6.25. Final probability maps provided by the binary-based (left) and concentration-based (right) approach using optimized parameters. The sources are located at (1, -4)m for acetone ('A') and (5, -4) for ethanol ('E'), depicted using letters.

In the cases with more than one source, since the algorithms assume that there is only one in the area, it does not make sense to take the probability at real source location and the distance to real source location as figures of merit. Only the entropy over the final maps is informative.

The statistics of the entropy for the experiments with two sources are summarized in the following boxplot:
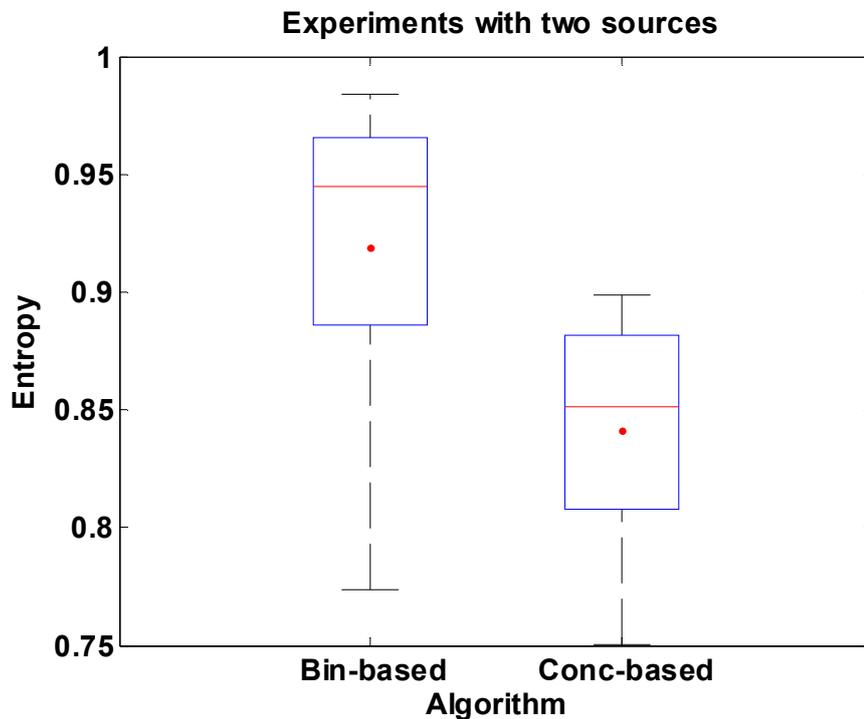
Figure 6.26. Entropy. Boxplot of the distribution of the results for the binary-based and concentration-based algorithms. The top and bottom of the box are the 25th and 75th percentiles. The point within the box is the mean and the horizontal line is the median. The whiskers extend to the most extreme data points which are not considered outliers.

It is seen in Figure 6.26 that the concentration-based approach provides more certainty in the location of the sources. The results from the Wilcoxon test suggest that the concentration-based approach performs significantly better in terms of entropy (p-value: 0.03).

In Figure 6.25 it is observed in the final probability maps that both source traces can be clearly identified in the concentration-based approach. However, these traces cannot be observed in the binary-based approach. In this last case, as the threshold level is increased, it is expected that all the readings from one of the sources will be discarded, then the algorithm only seeing the readings of the other source. Probably this threshold could be considered as optimal but then this becomes problematic if in fact the source that we want to detect is the weak one. On the other hand, in the concentration-based approach all the readings are computed, no information is discarded, and the algorithm finally can identify two clouds of probability in the map, which is advantageous.

6.8 Conclusions

In the present chapter, modifications of the previous work by Pang and Farrell (binary-based algorithm) have been introduced.

It is shown that the original algorithm by Pang and Farrell can be easily extended to work with multiple mobile sensors. All the information from the mobile sensors can be integrated in the algorithm, whatever their positions are. The algorithm only needs to know in which cell the concentration readings were obtained, and then a probability map will be recursively updated. Moreover, the mobile sensors do not need to solely perform plume tracking and might be used for other tasks at the same time as the most likely source position is being estimated using the available information found along the trajectory of the sensors.

Additionally, in a real scenario, pollution and some interfering substances may appear in the background, increasing the number of false alarms. Unlike the binary-based algorithm, which uses a threshold to assess whether a concentration is considered as a detection or non-detection event; our algorithm, based on continuous concentrations, builds a background model to assess whether a concentration comes from the background or from a source located further away. In addition, knowledge about the plume dispersion is introduced. This knowledge is based on previous work about plume dispersion in turbulent flow environments. Simulation results show that our algorithm behaves much more robustly in the presence of false alarms and better estimates the real source location.

All concentration readings are considered in our algorithm, incorporating them in a continuous manner instead of just using them as binary detections above a certain threshold. This fact removes the need for a threshold level, thereby improving the performance of the algorithm proposed by Pang and Farrell. Against to what might be expected, increasing the threshold in the binary-based approach does not always lead to an improvement in the algorithm's performance. There is a trade-off between the false alarm rejection and the false negatives (concentrations from the plume which are not detected). Increasing the threshold too much could lead to a very long sequence of non-detection events, making the source localization problem infeasible. This depends on the unknown background level; therefore the threshold would need to be set arbitrarily. Saying that the background is unknown is equivalent to saying that the source strength is unknown, because the background is low or high compared to the source strength, and it cannot be determined where the reading comes from.

In the concentration-based approach, considering all the concentration values reduces the number of false alarms (a background model is estimated) and reduces the

number of missed detections arising from setting the threshold too high since no threshold value is needed. However, the algorithm assumes that the source strength is known. Since this is a critical point in our proposal, a study of the robustness of the algorithm against deviations from the true value has been presented. It is shown that in cases where the source to be detected is weak (or the background is high) the concentration-based approach behaves more robustly if the source strength can be estimated to within at least two orders of magnitude. However, in cases where the source is very potent (or the background is low) using the binary-based algorithm and optimizing the threshold by some means or other might be adequate to solve the source localization problem; even though, as it has been shown, the results improve significantly using the concentration-based algorithm if the source strength can be estimated. Therefore, we consider that estimating the source strength instead of optimizing the threshold in the binary-based approach is a more promising direction for future research, since no information is dismissed when the concentration-based approach is used and it is thus more robust against false alarms and missed detections. Furthermore, the simulation results have been validated using real-world data acquired from a PID mounted on a mobile robot. In these results it is demonstrated that the concentration-based approach behaves as well as the optimum binary case; and that, unlike the binary-based approach, the algorithm could be suitable when more than one source is present in the exploration area, since no chemical information is discarded. However, tests have not been performed in the presence of high levels of interfering chemicals, which are the conditions that favour the most the present proposal.

6.9 References

Bakkum, EA & Duijm, NJ 2005, 'Vapour cloud dispersion', in *Methods for the calculation of physical effects 'Yellow Book'*, 3rd edn, The Hague, pp. 4.1-4.139.

Beychok, MR 2005, *Fundamentals of Stack Gas Dispersion*, Milton R. Beychok.

Crimaldi, JP, Wiley, MB & Koseff, JR 2002, 'The relationship between mean and instantaneous structure in turbulent passive scalar plumes', *Journal of Turbulences*, vol. 3, no. 14, pp. 1-24.

Fackrell, JE & Robins, AG 1982, 'Concentration fluctuations and fluxes in plumes from point sources in a turbulent boundary layer', *Journal of Fluid Mechanics*, vol. 117, pp. 1-26.

Farrell, JA, Murlis, J, Long, X, Li, W & Cardé, RT 2002, 'Filament-based atmospheric dispersion model to achieve short time-scale structure of odor plumes', *Environmental Fluid Mechanics*, vol. 2, pp. 143-169.

Farrell, JA, Pang, S & Li, W 2005, 'Chemical Plume Tracing via an Autonomous Underwater Vehicle', *IEEE Journal of Oceanic Engineering*, vol. 30, no. 2, pp. 428-442.

Gijbels, I & Hubert, M 2009, 'Robust and nonparametric statistical methods', in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, vol. 1, eds SD Brown, R Tauler & B Walczak, Elsevier, pp. 189-211.

Hahn, I, Brixey, LA, Wiener, RW, Henkle, SW & Baldauf, R 2009, 'Characterization of traffic-related PM concentration distribution and fluctuation patterns in near-highway urban residential street canyons', *Journal of Envorinmental monitoring*, vol. 11, pp. 2136-2145.

Jones, AR & Thomson, DJ 2006, 'Simulation of time series of concentration fluctuations in atmospheric dispersion using a correlation-distortion technique', *Boundary-Layer Meteorology*, vol. 118, pp. 25-54.

Li, J-G, Meng, Q-H, Wang, Y & Zeng, M 2011, 'Odor source localization using a mobile robot in outdoor airflow environments with a particle filter algorithm', *Autonomous Robots*, vol. 30, no. 3, pp. 281-292.

Martín, M, Cremades, LV & Santabàrbara, JM 1999, 'Analysis and modelling of time series of surface wind speed and direction', *International Journal of Climatology*, vol. 19, pp. 197-209.

Pang, S & Farrell, JA 2006, 'Chemical Plume Source Localization', *IEEE Transactions on systems, man, and cybernetics*, vol. 36, no. 5, pp. 1068-1080.

Papoulis, A 1984a, 'Chapter 8: Sequences of random variables', in *Probability, random variables, and stochastic processes*, McGraw-Hill, New York.

Papoulis, A 1984b, *Probability, random variables, and stochastic processes*, 2nd edn, McGraw-Hill, New York.

Soriano, C, Baldasano, JM, Buttler, WT & Moore, KR 2000, 'Circulatory patterns of air pollutants within the Barcelona air basin in a summertime situation: lidar and numerical approaches', *Boundary-Layer Meteorology*, vol. 98, pp. 33-55.

Sutton, OG 1947, 'The problem of diffusion in the lower atmosphere', *Q. J. R. Meteorol. Soc.*, vol. 73, pp. 257-281.

Turner, B 1994, *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*, CRC Press. Taylor and Francis group, North Carolina, USA

Webster, DR 2007, 'Structure of turbulent chemical plumes', in *Trace Chemical Sensing of Explosives*, ed. RL Woodfin, John Wiley & Sons, Inc, pp. 109-129.

Webster, DR & Weissburg, MJ 2001, 'Chemosensory guidance cues in a turbulent chemical odor plume', *Limnol. Oceanogr.*, vol. 46, no. 5, pp. 1034-1047.

Yee, E 2008, 'Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference', *Boundary-Layer Meteorology*, vol. 127, pp. 359-394.

Yee, E 2009, 'Probability law of concentration in plumes dispersing in an urban area', *Environmental Fluid Mechanics*, vol. 9, pp. 389-407.

Yee, E & Biltoft, CA 2004, 'Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles', *Boundary-Layer Meteorology*, vol. 111, pp. 363-415.

Yee, E & Chan, R 1997, 'A simple model for the probability density function of concentration fluctuations in atmospheric plumes', *Atmospheric Environment*, vol. 31, pp. 991-1002.

Yee, E, Chan, R, Kosteniuk, PR, Chandler, GM, Biltoft, CA & Bowers, JF 1994, 'Experimental measurements of concentration fluctuations and scales in a dispersing plume atmospheric surface-layer obtained using a very fast-response concentration detector', *Journal of Applied Meteorology*, vol. 33, no. 8, pp. 996-1016.

Yee, E, Chan, R, Kosteniuk, PR, Chandler, GM, Biltoft, CA & Bowers, JF 1995, 'The vertical structure of concentration fluctuation statistics in plumes dispersing in the atmospheric surface layer', *Boundary-Layer Meteorology*, vol. 76, no. 1-2, pp. 41-67.

Yee, E, Kosteniuk, PR, Chandler, GM, Biltoft, CA & Bowers, JF 1993, 'Statistical characteristics of concentration fluctuations in dispersing plumes in the atmospheric surface layer', *Boundary-Layer Meteorology*, vol. 65, no. 1-2, pp. 69-109.

Yee, E, Wang, B-C & Lien, F-S 2009, 'Probabilistic model for concentration fluctuations in compact-source plumes in an urbant environment', *Boundary-Layer Meteorology*, vol. 130, pp. 169-208.

Chapter 7

Conclusions of this Thesis

In the present thesis, signal processing approaches to improve the detection, identification, and quantitation of analytes when using mobility based analyzers have been presented. Applications using an Ion Mobility Spectrometer (IMS) and a Differential Mobility Analyzer (DMA) are shown.

We have demonstrated that standard multivariate data analysis tools are effective in DMA for explosive detection and identification and for identification and quantitation of volatile organic compounds (VOCs).

In chapter 4, a blind source separation method based on a modification of Multivariate curve resolution Alternating Least Squares (MCR-ALS) using L1 regularization LASSO (least absolute shrinkage and selection operator) is presented. It is shown that peaks in IMS spectra can be modelled as a dense superposition of Gaussian functions. The resulting bilinear decomposition is sparse and by virtue of the underlying peak model shape provides filtered spectra and also concentration signals.

We show that we can rely on the method in order to improve the detection of certain substances even in the presence of interfering chemicals. The provided results are based on synthetic generated data and real IMS data obtained from a continuous monitoring application such as a baggage security checkpoint.

As substance concentration increases, IMS spectra become more non-linear and more than one peak related to the target substance may appear. In chapter 5, a new methodology is proposed based on MCR-ALS followed by polynomial Partial Least Squares (poly-PLS) and it is shown capable of qualitative and quantitative analysis of IMS signals. Separate ionic concentration profiles and their corresponding spectral features are obtained from MCR-ALS, which provides a meaningful interpretation of the analysis.

Finally, it is shown even with partially selective sensors that the use of substance concentration can help to solve the source localization problem, especially in cases where the source strength is weak (or the background of interfering chemicals is high). Furthermore, we consider that integrating continuous concentration information

within the source localization algorithm provides a more promising direction for future research than managing responses of the sensors as detections or non-detections above a certain threshold. This is illustrated in chapter 6, where results from simulations and from real-world data are presented. Real experiments were carried out using a mobile robot mounting a photo ionization detector (PID) in indoors and outdoors environments under forced ventilation and turbulences. The obtained results validate the results obtained from simulations, thus the source localization algorithm developed in the thesis is effective in real scenarios.

Chapter 8

Resumen de la tesis

8.1 Introducción

A medida que la tecnología evoluciona, la instrumentación genera mayores volúmenes de datos, lo que dificulta que estos puedan ser tratados manualmente. Así que técnicas de procesado de señal y datos son necesarias con la finalidad de extraer la información relevante de las señales, que usualmente además son ruidosas.

La habilidad de detectar e identificar correctamente sustancias químicas se requiere en un gran número de aplicaciones, desde control medioambiental, aplicaciones biomédicas, evaluación de la calidad en comida y bebidas, exploración de áreas donde se han dispersado sustancias tóxicas, hasta operaciones humanitarias y de seguridad. En algunas de estas aplicaciones y escenarios, la muestra se toma manualmente y se analiza a posteriori en un laboratorio, así identificando y cuantificando los compuestos presentes; en otros en cambio, la localización de la fuente química es desconocida y por lo tanto las muestras no pueden tomarse manualmente. Es en este tipo de escenarios donde las sustancias deben ser identificadas y cuantificadas a distancia y además algoritmos de localización deben ser empleados para encontrar la posición de la fuente. Sin embargo, este problema está lejos de estar resuelto debido a la elevada complejidad que comporta la dispersión de agentes químicos en entornos reales (Kowadlo & Russell 2008). Además la presencia de interferentes dificulta las tareas de detección, cuantificación y localización.

Para la presente tesis nos centramos en sustancias químicas en fase gas y aplicaciones de campo las cuales requieren una monitorización continua.

8.1.1 Escenarios

De entre los distintos escenarios, algunos tienen implicaciones muy importantes por temas de seguridad. Por ejemplo, después de los atentados del 11-S (2001) en USA, hay una demanda creciente en cuanto a nuevas tecnologías sensoras complementarias al uso de perros entrenados para el escaneo de equipaje en aeropuertos para la detección de drogas y explosivos (Singh & Singh 2003; Moore 2007). Se precisa de instrumentación fiable para el análisis de volátiles, con gran sensibilidad (ppb o ppt) y selectividad y que requiera de poca preparación para la muestra. En la presente tesis, en el capítulo 4 y en (Pomareda et al. 2010), se presentan resultados empleando un

prototipo para el control de equipaje utilizando un espectrómetro de movilidad de iones (IMS) (Eiceman & Karpas 2005).

En aeropuertos, al escanear maletas, las muestras pueden tomarse manualmente y analizarse, en otros sin embargo la posición de la fuente puede ser desconocida. Por ejemplo, hay múltiples productos que se usan en la industria para distintas aplicaciones, como agentes químicos de guerra (CWA) o compuestos tóxicos industriales (TICs), pero que pueden producir efectos perniciosos para la salud en el caso de fugas incontroladas. Se ha demostrado la eficacia de múltiples tecnologías sensoras para la detección de estos compuestos (Sferopoulos 2009). No obstante, estas sustancias en manos de terroristas son un peligro para la sociedad. En un escenario urbano, es de importancia capital la detección de explosivos y sus precursores durante la fase de producción antes de que se produzca un atentado. Es posible la creación de una red sensores fijos y/o móviles parcialmente selectivos con la finalidad de muestrear el aire en busca de varias sustancias de interés. Las medidas de los sensores y su posición pueden ser enviadas a un sistema central donde se almacena y procesa toda la información con el objetivo de localizar posibles amenazas (Abbaspour & Mansouri 2005). Además, otra potencial aplicación de las redes de sensores es la monitorización ambiental, para controlar los niveles de contaminación.

Es importante remarcar que en escenarios de este estilo la localización de fuentes químicas (presumiblemente débiles) puede ser verdaderamente complicada debido además a condiciones meteorológicas cambiantes (atmósfera, temperatura, viento,...) o a cambios espaciales y temporales en las sustancias interferentes. Estos problemas son tratados en el capítulo 6.

8.1.2 Tecnologías de sensores de gas

Dependiendo del escenario, el abanico de sustancias químicas que deberían detectarse y/o cuantificarse puede ser muy amplio. Idealmente, uno querría tener tantos sensores específicos como sustancias a detectar. No obstante, esto en la práctica no siempre es posible, ya sea por precio, porque el sensor no existe o porque las sustancias a detectar son desconocidas. Como solución se propone el uso de sensores no selectivos los cuales responden a un gran número de sustancias diferentes y con un coste generalmente menor. Como contrapartida, sucede que sustancias interferentes enmascaran la señal de interés, lo cual dificulta la detección, cuantificación y localización de la fuente de volátiles. En la presente tesis se propone el uso de sensores de gas parcialmente selectivos para la detección y cuantificación de volátiles y explosivos mediante un procesado de señal adecuado (capítulos 3, 4 y 5).

Múltiples tecnologías pueden encontrarse en el mercado, de entre las cuales las más relevantes para la presente tesis doctoral son: detectores de fotoionización (PID)

(Daum et al. 2006), sensores de óxido metálico (MOX) (Meixner & Lampe 1996), espectrómetros de movilidad iónica (IMS) (Eiceman & Karpas 2005; Hill et al. 1990) y analizadores de movilidad diferencial (DMA) (Alonso et al. 2009; Pomareda et al. 2013). Típicamente, sensores parcialmente selectivos de distintos tipos (no solo MOXs) pueden agruparse en un mismo sistema con la finalidad de mejorar la selectividad. Esta es la filosofía seguida en las narices electrónicas (Persaud & Dodd 1982). En el caso de IMS y DMA, la mejora en la selectividad viene dada por la multidimensionalidad de la respuesta espectral y una buena resolución espectral; esto combinado con una mayor sensibilidad (menores límites de detección) y un decreciente coste hace que estos instrumentos sean cada vez más utilizados en un mayor número de aplicaciones.

8.1.2.1 Espectroscopía de movilidad de iones (IMS)

La tecnología IMS (Eiceman & Karpas 2005) es probablemente una de las más maduras para la detección de explosivos, CWA y TICs. Proporciona rápidos tiempos de análisis (1-5s), portabilidad, alta sensibilidad (ppb-ppt), no es necesario preparar la muestra y con un relativo bajo coste. El instrumento básico se compone de dos partes principales: la región reactiva y el tubo de derivas. La región reactiva incorpora una fuente de ionización, la cual determina la selectividad del instrumento y puede basarse en distintos principios de operación: radioactivo, lámpara UV (Baumbach et al. 2003), descarga de corona (Sabo, Matúška & Matejčík 2011), ESI (Hilton et al. 2010) o MALDI (Chen 2008). Los compuestos de la muestra son ionizados en esta región antes de entrar al tubo de derivas, el diseño del cual determina la sensibilidad y resolución del instrumento. Principalmente existen dos configuraciones para el tubo: lineal (DC) y asimétrico (AC). A grandes rasgos el funcionamiento es similar, los iones son separados dentro del tubo por masa, forma y carga al aplicarse un campo eléctrico (DC y/o AC) hasta llegar a un detector el cual recoge la carga de los iones generando un espectro de movilidad. En la presente tesis nos centramos en la configuración IMS típica (fuente radioactiva y tubo lineal).

Típicamente, el principal problema de IMS es la moderada selectividad. El problema se agrava cuando múltiples sustancias están presentes en la muestra y en presencia de interferentes (Hill & Simpson 1997; Márquez-Sillero et al. 2011) y humedad (Puton et al. 2012). Además, en la configuración típica, la respuesta del instrumento es no lineal al incrementarse la concentración de la sustancia y se pueden generar monómeros y dímeros dando como resultado múltiples picos en el espectro (Ewing, Eiceman & Stone 1999; Pomareda et al. 2012). Una manera de minimizar estos problemas es acoplar instrumentación adicional como paso previo a IMS para preseparar compuestos (Dworzanski et al. 1994; Kanu, Wu & Hill 2008). Otra solución complementaria consiste

en el uso de técnicas de análisis multivariante, lo cual se trata en los capítulos 4 y 5 y en (Pomareda et al. 2010; Pomareda et al. 2012).

8.1.2.2 Analizador de movilidad diferencial (DMA)

DMA es una configuración particular dentro de la familia IMS donde iones de distintas movilidades son separados en espacio en lugar de en tiempo de deriva. Una de las ventajas de esta configuración es que proporciona gran resolución lo cual implica una mayor selectividad, por contra el tiempo de análisis es un poco mayor. El DMA usado en la presente tesis ha sido desarrollado por (RAMEM) y se corresponde con la configuración de placas paralelas (Alonso et al. 2009; Santos et al. 2009). El instrumento se describe en (Pomareda et al. 2013), donde además se presentan resultados utilizando el DMA para detección y cuantificación de compuestos orgánicos volátiles (VOCs). Esto también se presenta en el capítulo 3, donde además se presentan resultados identificando varios explosivos.

8.1.3 Procesado de señal para espectros de IMS (IMS y DMA)

Para la detección y cuantificación de sustancias, la multidimensionalidad de los espectros de IMS y DMA hace que puedan ser utilizadas técnicas de procesado multivariante, las cuales proporcionan una mejora en la selectividad. Estas técnicas fueron anteriormente usadas para tratar datos procedentes de narices electrónicas o matrices de sensores (Scott, James & Ali 2006).

Sin embargo, los espectros de IMS son típicamente ruidosos, presentan una cierta línea de base y son sensibles a cambios en la humedad, la presión y la temperatura lo cual hace que la señal sea inestable, con desplazamiento de picos, lo cual conduce a desalineamientos de la señal; así que los espectros deben ser preprocesados antes de aplicar las técnicas multivariantes para detectar y cuantificar las sustancias de interés. Para la reducción del ruido pueden utilizarse filtros de mediana, de Savitzky-Golay (Savitzky & Golay 1964) o *wavelets* (Barclay, Bonner & Hamilton 1997). La línea de base puede ser sustraída mediante algoritmos iterativos (Komsta 2011; Zhang, Chen & Liang 2010; Gan, Ruan & Mo 2006) y algoritmos inspirados en *dynamic time warping* (DTW) (Ramaker et al. 2003; Tomasi, van den Berg & Andersson 2004) o *correlation optimized warping* (COW) (Tomasi, van den Berg & Andersson 2004; Skov et al. 2006) son útiles para corregir desplazamientos de los picos debidos a la humedad o la temperatura, por ejemplo *icoshift* (Savorani, Tomasi & Engelsen 2010).

Cuando se maneja gran cantidad de datos multidimensionales, mucha de la información contenida es redundante por lo que es útil reducir la dimensionalidad del espacio reteniendo solo la información esencial. Con esta finalidad existen diversas

técnicas, entre las cuales: redes neuronales, *principal component analysis* (PCA) (Wold, Esbensen & Geladi 1987), *partial least squares* (PLS) (Wold, Sjöström & Eriksson 2001), SIMPLISMA (Windig et al. 2005) o *multivariate curve resolution* (MCR) (Tauler, Kowalski & Fleming 1993).

Uno de los objetivos de la presente tesis es el de identificar y cuantificar sustancias en presencia de interferentes. Al usar métodos supervisados, modelos de entrenamiento pueden crearse en una fase previa, con muestras obtenidas en condiciones conocidas (supervisado), para luego clasificar automáticamente muestras desconocidas (asignación de clases) en la fase posterior de predicción. En esta última fase se requiere de un clasificador para asignar una clase a cada muestra desconocida. Típicamente el clasificador se aplica después de una etapa de reducción de dimensionalidad; para tal fin, pudiendo utilizarse alguno de los algoritmos anteriores y otras técnicas como *Linear discriminant analysis* (LDA) (Fisher 1936) o PLS-DA (Barker & Rayens 2003), las cuales incorporan información de las clases en el modelo para mejorar la clasificación de muestras nuevas. Como clasificador pueden utilizarse funciones lineales o no-lineales que proporcionan fronteras de decisión (Bishop 2006). Estas fronteras se generan en la etapa de entrenamiento a partir de los algoritmos mencionados anteriormente, por ejemplo. Una opción alternativa es usar el clasificador KNN, el cual es no-lineal y ha sido ampliamente utilizado en la literatura (Hastie, Tibshirani & Friedman 2003).

Cuando se analizan muestras de IMS (moderada selectividad y no linealidades) sin tener conocimiento acerca de la composición de la muestra, técnicas de separación ciega como PCA, SIMPLISMA o MCR proporcionan análisis cualitativos acerca de los componentes presentes en la muestra y la variación de su concentración con el tiempo. MCR-ALS (sección 1.4.2.3) es un método iterativo que puede inicializarse a partir de PCA o SIMPLISMA, y que permite la incorporación de conocimiento al modelo con el objetivo de mejorar el análisis de las muestras. Por ejemplo, puede imponerse que el resultado sea no negativo, que solo aparezca un pico por componente o que se conserve la carga del sistema, entre otros. En el capítulo 4 se demuestra que MCR es útil para interpretar el sistema en presencia de interferentes y en el capítulo 5 que es útil para analizar no linealidades causadas por variaciones en las concentraciones de las sustancias. Además de realizar una interpretación cualitativa de los resultados es importante cuantificar las sustancias presentes en las muestras (sección 1.4.2.4), para lo cual existen algoritmos robustos como PLS o poly-PLS (Wold, Kettaneh-Wold & Skagerberg 1989); este último útil para tratar con las no linealidades típicas en espectros de IMS. En el capítulo 5 muestras de IMS son cuantificadas, lo mismo se hace en el capítulo 3 con muestras de DMA.

8.1.4 Algoritmos de localización de fuentes químicas

Una vez tratado el problema de clasificar y cuantificar muestras desconocidas utilizando las técnicas mencionadas anteriormente, existe el problema de localizar la fuente química cuando ésta es desconocida. No obstante, la dispersión de volátiles es complicada en distintos entornos y el éxito del algoritmo de localización viene determinado por como éste está adaptado al entorno.

Principalmente, pueden diferenciarse dos grandes grupos: dispersión dominada por difusión y por turbulencias. Cuando los volátiles se dispersan debido a difusión, se forman distribuciones de gas con suaves gradientes de concentración; en el caso de dispersión por turbulencias, el gas forma meandros y se dispersa formando nubes de gas, existiendo así regiones intermitentes de alta concentración con gradientes muy abruptos en sus bordes, además estas regiones fluctúan en intensidad y dirección. Afortunadamente, la dispersión puede modelizarse mediante un modelo de pluma Gaussiana (GPM) (Bakkum & Duijm 2005). Sin embargo este modelo solo sirve para las características temporales promedio de la pluma. Para modelizar las características instantáneas (que serán las que los sensores medirán) se requieren modelos alternativos (Farrell et al. 2002). En la presente tesis, en el capítulo 6, se propone el uso de estadísticas basadas en la distribución clipped-gamma (CGD). Múltiples estudios empíricos demuestran que la CGD es apropiada para un gran rango de condiciones atmosféricas y a distintas distancias de la fuente (de metros a 1km) (Yee, Wang & Lien 2009; Yee 2009; Yee 2008; Yee & Biltoft 2004; Yee & Chan 1997).

Múltiples algoritmos de localización han sido propuestos en la literatura según las condiciones de flujo (difusión o turbulencias), donde se presentan resultados procedentes de simulaciones así como de experimentos en entornos reales utilizando robots móviles. Muchos de estos algoritmos han sido clasificados y recopilados en distintas *reviews* (Lilienthal, Loutfi & Duckett 2006; Kowadlo & Russell 2008; Mcgill & Taylor 2011; Ishida, Wada & Matsukura 2012).

Los primeros trabajos en el campo se realizaron asumiendo flujos dominados por difusión y suaves gradientes de concentración, guiando el robot hacia la dirección donde se maximizaba el gradiente (Rozas, Morales & Vega 1991). Esta técnica conocida como *chemotaxis* (Bell & Tobin 1982; Louis et al. 2008) fue posteriormente extendida a casos en los que los flujos eran turbulentos. En estos casos, distintos algoritmos empezaron a explotar también la información proporcionada por el flujo (*anemotaxis*), para guiar el robot hacia la fuente. Muchos de estos algoritmos fueron inspirados por el comportamiento de bacterias o animales como cangrejos, escarabajos, langostas, polillas, hormigas entre otros, los cuales son llamados

algoritmos bioinspirados. Todos estos algoritmos pueden incluirse en el grupo *reactive plume tracking*, el cual puede subdividirse en tres etapas: búsqueda de gas, rastreo de la pluma y declaración de la fuente (Lilienthal, Loutfi & Duckett 2006; Kowadlo & Russell 2008). Con la finalidad de mejorar la localización de la fuente, se implementan algoritmos que combinan *chemotaxis* y *anemotaxis*. Además, en la estrategia denominada *fluxotaxis* (Zarzhitsky, Spears & Spears 2005), se combina también la información del flujo de masa. Un enfoque alternativo denominado *infotaxis*, fue presentado en (Vergassola, Villermaux & Shraiman 2007), donde el robot debe decidir si moverse hacia la dirección donde se maximiza la ganancia de información o quedarse parado adquiriendo más datos para posteriormente ser explotados.

En la literatura además se demuestra que el uso simultáneo de múltiples robots móviles trabajando en equipo permite la implementación de algoritmos de localización más robustos (Mcgill & Taylor 2011). Sin embargo, algunos de los problemas en estos algoritmos de rastreo de plumas está en el hecho de que los robots están dedicados plenamente a la tarea de localizar la fuente y que debe decidirse cuando finalizar la búsqueda; este último problema solo se ha tratado en algunos trabajos (Lilienthal et al. 2004; Lilienthal et al. 2006; Loutfi & Coradeschi 2006).

No obstante, en algunas aplicaciones tan solo es necesario conocer la distribución de gas en el entorno o los robots deben ser destinados a otras tareas además de localizar la fuente. Para estas aplicaciones son apropiados los algoritmos de modelización de plumas (*plume modelling*). En estos algoritmos, la posición de la fuente es un parámetro más a estimar además de la distribución de gas, los cuales se actualizan en tiempo real a medida que nuevos datos están disponibles; así que el problema de decidir cuando finalizar la búsqueda desaparece. Información acerca de la dispersión de plumas se tiene en cuenta al implementar estos algoritmos ya sea mediante modelos analíticos (Ishida, Nakamoto & Moriizumi 1998; Marques, Nunes & Almeida 2002) o métodos probabilísticos/estocásticos (Pang & Farrell 2006). La integración de esta información adicional mejora la estimación de la posición de la fuente. Estos últimos métodos muestran un gran potencial, especialmente cuando se incorpora inferencia Bayesiana. Por ejemplo, en (Pang & Farrell 2006) se presenta un algoritmo eficiente el cual actualiza en tiempo real un mapa de probabilidad en la posición de la fuente integrando la secuencia de detecciones/no-detecciones a lo largo de la trayectoria del robot. Sin embargo, información importante de la concentración es descartada, lo cual puede ser crítico si quiere detectarse la fuente a grandes distancias. En algunos trabajos ya se sugiere que el utilizar la información de la concentración química puede ser útil para mejorar la localización de la fuente (Lo Iacono 2010). El tema de integrar todas las medidas dentro del algoritmo de localización no ha sido

tratado en profundidad en la literatura; por lo que es una de los objetivos de la tesis el demostrar que aporta una mejora (capítulo 6).

8.2 Técnicas Multivariantes aplicadas a un Analizador de Movilidad Diferencial (DMA) para la Detección de Explosivos e Identificación y Cuantificación de VOCs

En esta sección se expone un breve resumen del capítulo 3, donde varios compuestos orgánicos volátiles (VOCs) son identificados y cuantificados. También se identifican varios explosivos.

En la mayoría de aplicaciones, los instrumentos IMS y DMA solo se usan como detectores cualitativos por lo que solo interesa la ausencia o presencia de la sustancia de interés (*target substance*). La cuantificación de los analitos de interés suele ser típicamente univariante en tecnologías IMS, de modo que el área de un pico o su altura suele relacionarse con una concentración específica, pero esta solución no debería aplicarse cuando aparecen picos solapados debidos a interferencias. Tal y como se ha demostrado en los capítulos 4 y 5, el procesado de datos multivariante de espectros IMS permite minimizar los problemas debidos a una baja selectividad (interferencias cruzadas) así como detectar múltiples sustancias incluyendo interferencias. Además, con estas técnicas la cuantificación de las sustancias se mejora.

Aunque varios artículos han reportado el uso de técnicas de procesado de señal para análisis cualitativos (Rauch, Harrington & Davis 1998; Reese & Harrington 1999; Ochoa & Harrington 2005; Bota & Harrington 2006; Prasad et al. 2008; Pomareda et al. 2010; Karpas et al. 2012) o cuantitativos (Boger & Karpas 1994; Zheng, Harrington & Davis 1996; Fraga, Kerr & Atkinson 2009; Zamora & Blanco 2012; Pomareda et al. 2012) en datos de IMS, hasta donde sabemos, hay una falta de contribuciones analizando datos de instrumentos DMA usando técnicas multivariantes.

En el capítulo 3 se demuestra que la detección de explosivos y la identificación y cuantificación de VOCs con una estructura química similar, es viable mediante DMA. Por primera vez, se presentan resultados obtenidos aplicando técnicas multivariantes sobre espectros obtenidos usando un DMA de configuración original (RAMEM).

A pesar de que los resultados experimentales muestran como el sistema (DMA + procesado multivariante de datos) es capaz de discriminar correctamente entre distintos tipos de explosivos, el reducido número de muestras hace que tan solo puede considerarse este estudio como preliminar. Nuevas rondas de experimentos se

deberían llevar a cabo en el futuro con el objetivo de establecer el DMA de (RAMEM) como una herramienta efectiva para la detección de explosivos.

También se muestra como 4 VOCs (acetona, benceno, o-chileno y tolueno) han sido detectados, identificados y cuantificados usando el DMA. Se han determinado como límites de detección: 0.7ppm para acetona, 0.9ppm para benceno, 1.5ppm para o-chileno y 2.0ppm para tolueno. Los modelos PLS de calibración han sido capaces de predecir las concentraciones de los 4 analitos. El RMSEV (*root mean-squared error of validation*) para acetona, benceno, tolueno y o-chileno están en el rango entre 0.1ppm y 0.3ppm en validación. La identificación y clasificación de los compuestos está basada en PLS-DA y un clasificador KNN. El porcentaje global de clasificación fue del 77% con intervalos del (70-83)% al 95% de confianza.

Los resultados también muestran la gran influencia de la humedad. El próximo prototipo de DMA incluirá un condicionamiento del gas *sheath*, con elementos calefactores y un filtro para testar varios tamices moleculares. Esto reducirá la gran influencia de la humedad en los espectros de DMA, por lo que se espera que los resultados en general sean más reproducibles.

En el capítulo 3, además también se describe el procedimiento general para predecir (identificación y cuantificación) muestras desconocidas, en un entorno real en tiempo real. Primero, modelos de entrenamiento deben ser construidos usando datos obtenidos bajo condiciones controladas en el laboratorio; seguidamente, estos modelos serán usados en un entorno real con la finalidad de predecir las sustancias presentes y su concentración en tiempo real. A pesar de que el procedimiento se explica para el caso del DMA, es general y puede aplicarse a cualquier tipo de instrumentación que proporcione datos multidimensionales (matrices de sensores de gas, IMS o espectroscopía en general).

8.3 MCR-ALS *sparse* usando regularización L1 y Modelos de Forma de Pico Gaussiana

En esta sección se expone un breve resumen del capítulo 4.

Tal como se ha descrito en el capítulo 1, *Multivariate Curve Resolution Alternating Least Squares* (MCR-ALS) tiene como objetivo recuperar la evolución de la señales fuente (perfiles de concentración, en nuestro caso) y la matriz de mezclado (características espectrales) sin ningún paso previo de calibración supervisada. Adicionalmente, es bien conocido que imponer conocimiento adicional acerca de los métodos o los procesos, mediante restricciones (*constraints*) puede llevar a mejores soluciones y a interpretaciones de los resultados más fáciles ya que las ambigüedades

rotacionales (Jaumot & Tauler 2010) se minimizan y el espacio de las posibles soluciones se reduce. En este sentido, restringir la solución imponiendo varias *constraints* es una práctica habitual en MCR (Juan et al. 1997). Por ejemplo, algunas *constraints* blandas (Tauler 1995; Gemperline & Cash 2003; Bro & Sidiropoulos 1998) son: 1) número de componentes que se espera encontrar en la muestra, 2) no-negatividad, 3) unimodalidad, 4) selectividad y 5) clausura. Sin embargo, al tratar con señales de IMS, debido a su moderada selectividad (espectros complicados con múltiples picos y solapados), en un escenario real donde múltiples sustancias interferentes estarán presentes, es casi imposible el conocer de antemano si un espectro será unimodal o qué regiones son selectivas a uno o más iones específicos; no obstante, el integrar tanto conocimiento acerca del sistema como sea posible puede ser determinante para la detección de ciertos agentes químicos. Por lo tanto, en los experimentos que se presentan en el capítulo 4, tan solo se aplican las *constraints* 1 y 2 para condicionar las soluciones lo menos posible.

Mientras en su forma básica, MCR es una técnica blanda (*soft*) de modelización (no se impone ningún modelo subyacente), varios autores han propuesto versiones rígidas (*hard*) dónde modelos físico-químicos, los cuales caracterizan los procesos subyacentes, se imponen en la solución (Juan et al. 2000). Al usar estos modelos, se está incorporando conocimiento adicional, por lo que la solución está incluso más restringida (*hard constraint*) y las ambigüedades rotacionales se minimizan también.

Para algunas medidas espectroscópicas, los espectros se caracterizan por la presencia de series de picos. Para algunos instrumentos, existen modelos aproximados acerca de la forma de los picos (e.g. picos Gaussianos) (Felinger 1998). Sin embargo, en un escenario de separación ciega de fuentes (*Blind source separation*), ni el número de picos ni su posición son conocidos de antemano. Este hecho dificulta seriamente la aplicación directa de modelos para la forma de los picos dentro del bucle de *alternating least squares* (ALS) de MCR.

En esta tesis se propone modelizar los espectros fuente como una densa superposición de picos Gaussianos y aplicar una regularización basada en norma L1 (L1-*norm*) para obtener una solución *sparse* y así extraer automáticamente el número correcto de picos y su posición sin haber impuesto a priori ni su posición ni su número. Para conseguir esto, se introduce la técnica LASSO (*Least Absolute Shrinkage and Selection Operator*) dentro del bucle MCR-ALS para modelizar los espectros. El nuevo algoritmo se ha denominado MCR-LASSO. LASSO fue propuesto por Tibshirani (Tibshirani 1996) y se conoce como *basis pursuit* (Chen, Donoho & Saunders 1998) o *compressed sensing* (Donoho 2006) en el campo de procesado de señal. Para testar este concepto se ha aplicado el MCR-ALS básico y MCR-LASSO a señales sintéticas y reales. Las señales reales corresponden a espectros de IMS obtenidos usando un prototipo para el

escaneo de equipaje, el cual imita un control de seguridad de aeropuerto.

Los resultados de experimentos sintéticos demuestran que en condiciones exigentes (alto nivel de ruido, perfiles de concentración similares, espectros solapados, y picos asimétricos) MCR-LASSO proporciona mejores estimaciones en la evolución temporal y los espectros de las componentes subyacentes. Se ha demostrado que la densa superposición de Gaussianas es capaz de modelizar picos asimétricos más anchos de los habitualmente encontrados en espectroscopía.

Por otro lado, se ha demostrado que MCR-LASSO proporciona mejor resolución en dos experimentos reales usando el citado prototipo de escaneo de equipaje. En el primer experimento se muestra que las no-linealidades típicas de IMS pueden ser tratadas mediante la introducción de componentes adicionales en el sistema. El segundo experimento presenta una mezcla más complicada y la presencia de agentes químicos interferentes. Los resultados obtenidos de MCR-LASSO contienen un nivel de ruido inferior, no solo en los espectros sino también en los perfiles de concentración, lo cual provoca que la interpretación cualitativa de los resultados resulte más sencilla si lo comparamos con los resultados obtenidos a partir de MCR-ALS y SIMPLISMA. Es importante remarcar que el método basado en LASSO puede usarse en combinación con otros modelos de picos aparte del modelo Gaussiano escogido para la presente tesis.

El uso de una solución regularizada a partir de la norma L1 permite usar un modelo flexible (en este caso una densa superposición de picos Gaussianos) que de otra manera derivaría en un problema de mínimos cuadrados muy mal condicionado, y aún así encontrar una solución *sparse* de complejidad limitada.

A pesar de que el trabajo realizado se ha basado en espectros de IMS y picos Gaussianos, MCR-LASSO puede aplicarse a otras configuraciones dado que los modelos espectrales pueden basarse en una densa superposición lineal de regresores.

8.4 Análisis cualitativo y Predicción cuantitativa de Espectros IMS no lineales

En esta sección se expone un breve resumen del capítulo 5, donde se propone una nueva metodología para el análisis de series temporales de espectros de IMS.

El método propuesto combina las ventajas de MCR-ALS para una interpretación óptima de las características físicas y químicas del sistema (información cualitativa), y de una técnica de calibración multivariante como *polynomial partial least squares* (poly-PLS) (Wold, Kettaneh-Wold & Skagerberg 1989; Rosipal 2008) para una mejor cuantificación (información cuantitativa) de nuevas muestras. Los datos analizados presentan un

comportamiento fuertemente no-lineal a medida que se incrementa la concentración de la sustancia.

Se ha demostrado en esta tesis que MCR-ALS es un método adecuado para el estudio de datos de IMS de segundo orden. Usando SIMPLISMA y MCR-ALS, los espectros de IMS se resuelven en componentes puras y se obtiene una estimación cualitativa para los espectros y los perfiles de concentración de estas componentes. A pesar de que estas técnicas se basan en una decomposición bilineal ($D = C \cdot S^T + E$), comportamiento no-lineal puede ser modelizado añadiendo más componentes al modelo. Después del preprocesado de los espectros, MCR-ALS se aplica con la finalidad de obtener información acerca de las especies iónicas que aparecen en el tubo de deriva de IMS y su evolución con la concentración del analito. Resolviendo la matriz de datos, se obtienen los perfiles de concentración y los espectros puros de las diferentes especies iónicas.

A pesar de que los monómeros y los dímeros pertenezcan a la misma sustancia, sus picos son modelizados como componentes separadas en los analitos estudiados (2-butanona y etanol). Por lo tanto, MCR-ALS proporciona una buena manera para obtener una interpretación física y química del sistema incluso en presencia de comportamientos fuertemente no-lineales. A pesar de que esta información cualitativa no puede usarse directamente para realizar una calibración, puede ser usada de manera efectiva por métodos de calibración multivariante tales como poly-PLS para construir modelos de calibración. Se usa poly-PLS con la finalidad de construir modelos de calibración usando los perfiles de concentración que se obtienen de MCR-ALS. Los resultados obtenidos muestran que se consiguen predicciones similares o mejores comparado con otras metodologías estándar de calibración univariante y multivariante. Así que, se demuestra que esta nueva metodología puede ser útil en casos en los cuales el comportamiento de las intensidades de los picos es no-lineales a medida que se incrementa la concentración de la sustancia.

Para las matrices de datos estudiadas, los resultados cualitativos muestran como las técnicas estándar de calibración multivariante funcionan en general mejor que las técnicas univariantes, especialmente cuando picos en los espectros aparecen solapados. Las técnicas multivariantes son capaces de modelizar comportamientos no-lineales a partir de la adición de más componentes al modelo. Los datos presentados incluyen comportamientos fuertemente no-lineales al incrementar la concentración de la sustancia. Mientras que PLS es capaz de manejar comportamientos ligeramente no-lineales, comportamientos fuertemente no-lineales son manejados de manera más eficiente por métodos como poly-PLS. A pesar de que la capacidad predictiva es similar, los resultados obtenidos son usualmente difíciles de interpretar usando estas técnicas de calibración, ya que, con el objetivo de modelizar no-linealidades, el número

de variables latentes en el modelo es típicamente mayor que el número de picos. Si se usa MCR-ALS como paso previo a la calibración, la interpretación cualitativa de los resultados es más sencilla y además el número de variables latentes queda fijado, de esta manera reduciendo la complejidad del modelo de calibración y reduciendo el número de parámetros a optimizar mediante validación cruzada.

Además, se han mostrado los resultados obtenidos a partir de réplicas. Estos resultados sugieren que los modelos de calibración construidos en un día determinado no pueden transferirse a otros días. Este punto queda abierto para trabajo futuro.

8.5 Localización de Fuentes Químicas usando Múltiples Sensores Móviles mediante Inferencia Bayesiana en presencia de Niveles Químicos Interferentes

Los experimentos de navegación hacia fuentes químicas están fuertemente limitados por el rápido decaimiento de la concentración de la fuente de interés al incrementarse la distancia. Límites de detección pobres hacen que la pluma solo puede ser detectada de manera efectiva en un área reducida. Esto es especialmente importante en aplicaciones en las cuales la zona de búsqueda tenga una extensión de cientos de miles de metros cuadrados. A no ser que la fuente sea muy potente, la tarea de localización de fuentes químicas resulta prácticamente imposible. En tales casos, debería situarse el umbral de detección muy cercano al nivel de ruido, pero esto generaría un gran número de falsas alarmas lo que conduciría a que la mayoría de algoritmos de localización fallaran de manera catastrófica. Por lo que se sabe, ninguno de los métodos públicos trata este problema.

Por otro lado, sensores químicos de bajo coste e incluso detectores de gama media (e.g. espectrómetros de movilidad iónica) tienen tan solo una selectividad limitada. En un escenario real, pueden existir señales de fondo (*background*) debidas a una multitud de sustancias químicas procedentes de la contaminación ambiental. Debido a la limitada selectividad de los detectores, habrá sustancias que generarán interferencias en las lecturas del detector, dificultando las tareas de detección y localización de la sustancia de interés. La combinación de ruido electrónico en el detector y principalmente sustancias interferentes resultará en lecturas remanentes los cuales pueden cambiar con el tiempo y la posición del detector. Estas derivas en los niveles de *background* dificultan la selección del umbral de detección óptimo. Por lo que se sabe, este problema no ha sido abordado previamente en la literatura.

Con el objetivo de tratar estos problemas, los algoritmos probabilísticos como los métodos Bayesianos de mapeo de pluma parecer ser una buena elección. Tales técnicas han resultado ser efectivas en muchas áreas de la robótica y proporcionan un buen punto de inicio para tratar los problemas. Pang y Farrell publicaron un algoritmo de mapeo de la verosimilitud en la posición de la fuente basado en inferencia Bayesiana (Pang & Farrell 2006). La principal idea consiste en implementar una aproximación estocástica (Farrell et al. 2002; Papoulis 1984) para la modelización de la pluma y en estimar la posición más probable de la fuente teniendo en cuenta la secuencia de eventos de detección/no-detección y las medidas del flujo a lo largo de la trayectoria del robot. A pesar que en el trabajo original, se presentan resultados realizando un rastreo de la pluma, tal rastreo no es estrictamente necesario para obtener una buena estimación de la posición de la fuente, con lo que los robots podrían ser destinados a otras tareas además de localizar la fuente. Actualizando recursivamente un mapa de probabilidad mediante inferencia Bayesiana se estima la posición más probable para la fuente.

Sin embargo, el citado algoritmo tan solo considera el uso de eventos de detección binarios. En otras palabras, no se usa la información procedente de las concentraciones químicas medidas por los sensores para construir el mapa de probabilidad. El algoritmo tan solo considera concentraciones por encima de un cierto umbral como eventos de detección o no-detección. Además, después de situar el umbral a un cierto nivel, se asume que la tasa de falsas alarmas es muy baja; no obstante, tal asunción no puede tomarse como cierta en un escenario real donde señales interferentes estarán presentes, a no ser que el umbral se sitúe a un nivel muy elevado, pero esta opción reducirá drásticamente la distancia máxima a la que podrá detectarse la fuente. Por lo tanto, existe un compromiso; por un lado, el umbral tiene que situarse lo suficientemente bajo (cerca del límite de detección) si la fuente quiere detectarse a grandes distancias; por otro lado, tiene que ser lo suficientemente alto como para minimizar la aparición de falsas alarmas. Por lo que, como situar el umbral es un tema crítico en entornos reales al utilizar las estrategias actuales, especialmente cuando no se tiene ninguna información acerca del background o de la potencia de emisión de la fuente.

Una de las motivaciones para la presente tesis es la de mejorar el algoritmo de localización de plumas basado en inferencia Bayesiana previamente descrito por Pang y Farrell (Pang & Farrell 2006), usando la información de las concentraciones químicas y extendiéndolo a entornos reales donde señales interferentes pueden aparecer. En nuestra propuesta, el algoritmo evalúa la probabilidad posterior que una cierta concentración química proceda del *background* o de una fuente química emitiendo continuamente a mayor distancia. Este enfoque elimina la necesidad de utilizar un

umbral. Así que, se reformula el algoritmo de Pang para ser usado con concentraciones continuas en lugar de simples eventos de detección binarios. Además, el algoritmo se extiende para trabajar con múltiples robots móviles, integrando simultáneamente sus lecturas sin importar de donde proceden. Esta nueva propuesta requiere un modelo probabilístico para el *background* y para la pluma, los cuales se han descrito en el capítulo 6. Estos modelos están basados en trabajos previos acerca de dispersión de plumas en entornos con flujos turbulentos (Yee, Wang & Lien 2009; Yee 2009; Yee 2008; Yee & Biltoft 2004; Yee & Chan 1997).

Los resultados obtenidos mediante simulaciones realistas muestran que el algoritmo desarrollado durante la tesis se comporta de manera mucho más robusta que el algoritmo de Pang en presencia en falsas alarmas y que la estimación en la posición real de la fuente química es más exacta. A pesar de lo que uno podría esperar inicialmente, no siempre el incrementar el umbral en el algoritmo de Pang conduce a una mejora. Situar el umbral demasiado alto conduce a situaciones en las cuales se generan largas secuencias de eventos de no-detección lo que provoca que el problema de localización de fuentes resulte inviable. Esto además depende del nivel de *background*, el cual es desconocido, por lo que en principio el umbral debe situarse a un nivel arbitrario. El decir que el nivel de background es desconocido es equivalente a afirmar que la potencia de emisión de la fuente también lo es, ya que el background es alto o bajo relativo a la potencia de la fuente.

En nuestra propuesta, el hecho de utilizar todas las medidas procedentes de los sensores reduce el número de falsas alarmas (se modeliza el *background*) y reduce el número de falsos negativos ya que no existe ningún umbral. Sin embargo, el algoritmo asume que la potencia de emisión de la fuente es conocida. Ya que esto es un punto crítico en nuestra propuesta, se ha realizado un estudio de la robustez del algoritmo contra desviaciones respecto al valor real. En los resultados se muestra que en los casos en los que la fuente es débil (o *background* alto), nuestra estrategia se comporta de manera mucho más robusta si la potencia de la fuente se puede estimar dentro de un rango de dos órdenes de magnitud alrededor del valor real. Sin embargo, en los casos en los que la fuente es potente (o *background* bajo), el algoritmo de Pang puede ser suficiente para resolver el problema de localización de la fuente (optimizando el umbral de alguna manera), a pesar de que, tal como se muestra en los resultados, nuestro algoritmo proporciona mucho mejores resultados si la potencia de emisión puede estimarse. Por lo tanto, consideramos que estimar la potencia de emisión de la fuente ofrece una mejor dirección para trabajo futuro que no el hecho de intentar encontrar un método para optimizar el umbral en el algoritmo de Pang. Además, la

potencia de emisión de la fuente es un parámetro de directo interés para un usuario a diferencia del umbral de detección.

En el capítulo 6 también se validan los resultados obtenidos de las simulaciones a partir de datos reales obtenidos de experimentos realizados en el *Mobile Robotics and Olfaction Lab* (Örebro, Suecia). Veinte experimentos fueron llevados a cabo utilizando un robot móvil el cual incorporaba un detector PID y un anemómetro ultrasónico, en condiciones de flujo forzado turbulento, en interiores y exteriores, con obstáculos y sin obstáculos y con una o dos fuentes. El robot se programó para explorar el área aleatoriamente parando durante 30s en distintas posiciones. Los resultados muestran que el algoritmo desarrollado durante la tesis se comporta de manera más robusta en promedio en términos de: 1) distancia del máximo de probabilidad a la posición real de la fuente, 2) probabilidad en la posición de la fuente, 3) certidumbre en la posición de la fuente medida mediante la entropía; y que, a diferencia del algoritmo de Pang y Farrell, nuestra propuesta puede ser adecuada cuando más de una fuente está presente en el área de exploración al observarse nubes de probabilidad alrededor de las posiciones de las fuentes.

8.6 Conclusiones

En la presente tesis, se han presentado técnicas de procesado de señal para mejorar la detección, identificación y cuantificación de analitos usando analizadores basados en movilidad iónica (IMS y DMA); para los cuales se han presentado algunas aplicaciones.

Además, se ha demostrado que la información de la concentración química, obtenida después de procesar los datos de los sensores, puede integrarse de manera efectiva en un algoritmo de localización de fuentes de gas con el objetivo de mejorar la localización de la fuente.

En el capítulo 3 se muestra como técnicas estándar de análisis multivariante se aplican de manera efectiva a la tecnología DMA para la detección de explosivos y la identificación y cuantificación de VOCs.

En el capítulo 4, se propone un nuevo método basado en MCR-ALS y LASSO (*least absolute shrinkage and selection operator*). Se muestra que los picos de los espectros de IMS se pueden modelizar como una superposición de funciones Gaussianas. Mostramos que podemos usar el método para mejorar la detección de ciertas sustancias incluso en presencia de agentes químicos interferentes. Los resultados están basados en datos generados de manera sintética y en datos de IMS obtenidos a partir de una aplicación de monitorización continua, tal como un portal de seguridad para el control de equipaje.

A medida que la concentración de la sustancia aumenta, los espectros de IMS son cada vez más no-lineales y es posible que aparezcan más picos relacionados con la sustancia de interés. En el capítulo 5 se propone una nueva metodología basada en MCR-ALS y poly-PLS para mejorar los análisis cualitativo y cuantitativo de las señales de IMS. Los perfiles de concentración y las características espectrales se obtienen a partir de MCR-ALS, método el cual proporciona una buena interpretación de las distintas especies iónicas con total sentido físico y químico.

A pesar de que las técnicas de procesado de señal utilizadas durante la tesis han sido testadas en condiciones de laboratorio, se muestra que son útiles para construir modelos de calibración, los cuales podrán ser usados en escenarios reales para predecir muestras desconocidas (identificación y cuantificación).

A pesar de utilizar sensores no-selectivos en un entorno abierto, se muestra que el uso de la información continua de la concentración química puede ayudar a resolver el problema de localización de fuentes químicas, especialmente en esos casos en los que la potencia de emisión de la fuente sea débil. Además, consideramos que la integración de la concentración química dentro del algoritmo proporciona dirección más prometedora para trabajo futuro que no el simple hecho de tratar los sensores como detectores binarios situando un umbral a su respuesta. Todo esto se ilustra en el capítulo 6, donde se presentan los resultados obtenidos mediante simulaciones realistas y datos reales.

Se llevó a cabo una serie de experimentos reales utilizando un robot móvil que incorporaba un detector de fotoionización (PID). Los experimentos se realizaron en interiores y exteriores bajo condiciones de ventilación forzada y flujo turbulento. Los resultados obtenidos validan aquellos obtenidos mediante las simulaciones; así que se muestra la efectividad del algoritmo de localización desarrollado durante la tesis operando en entornos reales.

8.7 Referencias

Abbaspour, M & Mansouri, N 2005, 'City hazardous gas monitoring network', *Journal of Loss Prevention in the Process Industries*, vol. 18, no. 4-6, pp. 481-487.

Alonso, M, Santos, JP, Hontanon, E & Ramiro, E 2009, 'First Differential Mobility Analysis (DMA) Measurements of Air Ions Produced by Radioactive Source and Corona', *Aerosol and Air Quality Research*, vol. 9, no. 4, pp. 453-457.

Bakkum, EA & Duijm, NJ 2005, 'Vapour cloud dispersion', in *Methods for the calculation of physical effects 'Yellow Book'*, 3rd edn, The Hague, pp. 4.1-4.139.

Barclay, VJ, Bonner, RF & Hamilton, IP 1997, 'Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression', *Analytical Chemistry*, vol. 69, no. 1, pp. 78-90.

Barker, M & Rayens, W 2003, 'Partial least squares for discrimination', *Journal of Chemometrics*, vol. 17, no. 3, pp. 166-173.

Baumbach, JI, Sielemann, S, Xie, ZY & 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*, vol. 75, no. 6, pp. 1483-1490.

Bell, WJ & Tobin, TR 1982, 'Chemo-Orientation', *Biological Reviews*, vol. 57, no. 2, pp. 219-260.

Bishop, CM 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc.

Boger, Z & Karpas, Z 1994, 'Use of neural networks for quantitative measurements in ion mobility spectrometry (IMS)', *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 3, pp. 576-580.

Bota, GM & Harrington, PB 2006, 'Direct detection of trimethylamine in meat food products using ion mobility spectrometry', *Talanta*, vol. 68, no. 3, pp. 629-635.

Bro, R & Sidiropoulos, ND 1998, 'Least Squares algorithms under unimodality and non-negativity constraints', *Journal of Chemometrics*, vol. 12, pp. 223-247.

Chen, P 2008, Applications of Chemometric Algorithms to Ion Mobility Spectrometry and Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry, thesis, Ohio University.

Chen, SS, Donoho, D & Saunders, M 1998, 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33-61.

Daum, KA, Watrous, MG, Neptune, MD, Michael, DI, Hull, KJ & Evans, JD 2006, *Data for First Responder Use of Photoionization Detectors for Vapor Chemical Constituents*, INL/EXT-05-00165; TRN: US200801%%351.

Donoho, D 2006, 'Compressed sensing', *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306.

Dworzanski, JP, Kim, MG, Snyder, AP, Arnold, NS & Meuzelaar, HLC 1994, 'Performance Advances in Ion Mobility Spectrometry through combination with high-speed vapor sampling, preconcentration and separation techniques', *Analytica Chimica Acta*, vol. 293, no. 3, pp. 219-235.

Eiceman, GA & Karpas, Z 2005, *Ion Mobility Spectrometry*, second edn, CRC Taylor and Francis Group, Boca Raton.

Ewing, RG, Eiceman, GA & Stone, JA 1999, 'Proton-bound cluster ions in ion mobility spectrometry', *International Journal of Mass Spectrometry*, vol. 193, no. 1, pp. 57-68.

Farrell, JA, Murlis, J, Long, X, Li, W & Cardé, RT 2002, 'Filament-based atmospheric dispersion model to achieve short time-scale structure of odor plumes', *Environmental Fluid Mechanics*, vol. 2, pp. 143-169.

Felinger, A 1998, 'Peak Shape analysis', in *Data analysis and signal processing in chromatography*, vol. 21, Elsevier Science, pp. 97-124.

Fisher, RA 1936, 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics*, vol. 7, pp. 179-188.

Fraga, CG, Kerr, DR & Atkinson, DA 2009, 'Improved quantitative analysis of ion mobility spectrometry by chemometric multivariate calibration', *Analyst*, vol. 134, no. 11, pp. 2329-2337.

Gan, F, Ruan, G & Mo, J 2006, 'Baseline correction by improved iterative polynomial fitting with automatic threshold', *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 59-65.

Gemperline, PJ & Cash, E 2003, 'Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions', *Analytical Chemistry*, vol. 75, no. 16, pp. 4236-4243.

Hastie, T, Tibshirani, R & Friedman, JH 2003, *The Elements of Statistical Learning*, Springer.

Hilton, CK, Krueger, CA, Midey, AJ, Osgood, M, Wu, J & Wu, C 2010, 'Improved analysis of explosives samples with electrospray ionization-high resolution ion mobility spectrometry (ESI-HRIMS)', *International Journal of Mass Spectrometry*, vol. 298, pp. 64-71.

Hill, HH, Siems, WF, Stlouis, RH & McMinn, DG 1990, 'Ion Mobility Spectrometry', *Analytical Chemistry*, vol. 62, no. 23, pp. A1201-A1209.

Hill, HH & Simpson, G 1997, 'Capabilities and Limitations of Ion Mobility Spectrometry for Field Screening Applications', *Annual Review of Physical Chemistry*, vol. 49, pp. 203-232.

Ishida, H, Nakamoto, T & Moriizumi, T 1998, 'Remote sensing of gas/odor source location and concentration distribution using mobile system', *Sensors and Actuators B: Chemical*, vol. 49, no. 1-2, pp. 52-57.

Ishida, H, Wada, Y & Matsukura, H 2012, 'Chemical Sensing in Robotic Applications: A Review', *Sensors Journal, IEEE*, vol. 12, no. 11, pp. 3163-3173.

Jaumot, J & Tauler, R 2010, 'MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution', *Chemometrics and Intelligent Laboratory Systems*, vol. 103, no. 2, pp. 96-107.

Juan, Ad, Maeder, M, Martinez, M & Tauler, R 2000, 'Combining hard and soft modelling to solve kinetic problems', *Chemometrics and Intelligent Laboratory Systems*, vol. 54, pp. 123-141.

Juan, AD, VanderHeyden, Y, Tauler, R & Massart, DL 1997, 'Assessment of new constraints applied to the alternating least squares method', *Analytica Chimica Acta*, vol. 346, pp. 307-318.

Kanu, AB, Wu, C & Hill, HH 2008, 'Rapid preseparation of interferences for ion mobility spectrometry', *Analytica Chimica Acta*, vol. 610, no. 1, pp. 125-134.

Karpas, Z, Guamán, AV, Calvo, D, Pardo, A & Marco, S 2012, 'The potential of ion mobility spectrometry (IMS) for detection of 2,4,6-trichloroanisole (2,4,6-TCA) in wine', *Talanta*, vol. 93, no. 0, pp. 200-205.

Komsta, L 2011, 'Comparison of Several Methods of Chromatographic Baseline Removal with a New Approach Based on Quantile Regression', *Chromatographia*, vol. 73, no. 7-8, pp. 721-731.

Kowadlo, G & Russell, RA 2008, 'Robot Odor Localization: A Taxonomy and Survey', *Int. J. Rob. Res.*, vol. 27, no. 8, pp. 869-894.

Lilienthal, A, Loutfi, A & Duckett, T 2006, 'Airborne Chemical Sensing with Mobile Robots', *Sensors*, vol. 6, no. 11, pp. 1616-1678.

Lilienthal, A, Ulmer, H, Frohlich, H, Werner, F & Zell, A 2004, 'Learning to detect proximity to a gas source with a mobile robot', in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, pp. 1444-1449 vol.2.

Lilienthal, AJ, Duckett, T, Ishida, H & Werner, F 2006, 'Indicators of Gas Source Proximity using Metal Oxide Sensors in a Turbulent Environment', in *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, pp. 733-738.

Lo Iacono, G 2010, 'A Comparison of Different Searching Strategies to Locate Sources of Odor in Turbulent Flows', *Adaptive Behavior*, vol. 18, no. 2, pp. 155-170.

Louis, M, Huber, T, Benton, R, Sakmar, TP & Vosshall, LB 2008, 'Bilateral olfactory sensory input enhances chemotaxis behavior', *Nat Neurosci*, vol. 11, no. 2, pp. 187-199.

Loutfi, A & Coradeschi, S 2006, 'Smell, think and act: A cognitive robot discriminating odours', *Autonomous Robots*, vol. 20, no. 3, pp. 239-249.

Marques, L, Nunes, U & Almeida, ATd 2002, 'Olfaction-based mobile robot navigation', *Thin Solid Films*, vol. 418, pp. 51-58.

Márquez-Sillero, I, Aguilera-Herrador, E, Cárdenas, S & Valcárcel, M 2011, 'Ion-mobility spectrometry for environmental analysis', *TrAC Trends in Analytical Chemistry*, vol. 30, no. 5, pp. 677-690.

Mcgill, K & Taylor, S 2011, 'Robot algorithms for localization of multiple emission sources', *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1-25.

Meixner, H & Lampe, U 1996, 'Metal oxide sensors', *Sensors and Actuators B: Chemical*, vol. 33, no. 1-3, pp. 198-202.

Moore, D 2007, 'Recent Advances in Trace Explosives Detection Instrumentation', *Sensing and Imaging: An International Journal*, vol. 8, no. 1, pp. 9-38.

Ochoa, ML & Harrington, PB 2005, 'Chemometric studies for the characterization and differentiation of microorganisms using in situ derivatization and thermal desorption ion mobility spectrometry', *Analytical Chemistry*, vol. 77, no. 3, pp. 854-863.

Pang, S & Farrell, JA 2006, 'Chemical Plume Source Localization', *IEEE Transactions on systems, man, and cybernetics*, vol. 36, no. 5, pp. 1068-1080.

Papoulis, A 1984, *Probability, random variables, and stochastic processes*, 2nd edn, McGraw-Hill, New York.

Persaud, K & Dodd, G 1982, 'Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose', *Nature*, vol. 299, no. 5881, pp. 352-355.

Pomareda, V, Calvo, D, Pardo, A & Marco, S 2010, 'Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra', *Chemometrics and Intelligent Laboratory Systems*, vol. 104, no. 2, pp. 318-332.

Pomareda, V, Guamán, AV, Mohammadnejad, M, Calvo, D, Pardo, A & Marco, S 2012, 'Multivariate curve resolution of nonlinear ion mobility spectra followed by multivariate nonlinear calibration for quantitative prediction', *Chemometrics and Intelligent Laboratory Systems*, vol. 118, no. 0, pp. 219-229.

Pomareda, V, Lopez-Vidal, S, Calvo, D, Pardo, A & Marco, S 2013, 'A novel differential mobility analyzer as a VOC detector and multivariate techniques for identification and quantification', *Analyst*, no. 138, pp. 3512-3521.

Prasad, S, Pierce, KM, Schmidt, H, Rao, JV, Guth, R, Synovec, RE, Smith, GB & Eiceman, GA 2008, 'Constituents with independence from growth temperature for bacteria using pyrolysis-gas chromatography/differential mobility spectrometry with analysis of variance and principal component analysis', *Analyst*, vol. 133, no. 6, pp. 760-767.

Puton, J, Holopainen, SI, Makinen, MA & Sillanpaa, MET 2012, 'Quantitative Response of IMS Detector for Mixtures Containing Two Active Components', *Analytical Chemistry*, vol. 84, no. 21, pp. 9131-9138.

Ramaker, HJ, van Sprang, ENM, Westerhuis, JA & Smilde, AK 2003, 'Dynamic time warping of spectroscopic BATCH data', *Analytica Chimica Acta*, vol. 498, no. 1-2, pp. 133-153.

RAMEM, *http://www.ioner.eu/*, RAMEM S.A.

Rauch, PJ, Harrington, PD & Davis, DM 1998, 'Chemometric resolution of mixture components by cleardown rates', *Analytical Chemistry*, vol. 70, no. 4, pp. 716-723.

Reese, ES & Harrington, PD 1999, 'The analysis of methamphetamine hydrochloride by thermal desorption ion mobility spectrometry and SIMPLISMA', *Journal of Forensic Sciences*, vol. 44, no. 1, pp. 68-76.

Rosipal, R 2008, 'Nonlinear Partial Least Squares An Overview', in *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pp. 169-189.

Rozas, R, Morales, J & Vega, D 1991, 'Artificial smell detection for robotic navigation', in *Advanced Robotics, 1991. 'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*, pp. 1730-1733 vol.2.

Sabo, M, Matúška, J & Matejčík, Š 2011, 'Specific O2 generation in corona discharge for ion mobility spectrometry', *Talanta*, vol. 85, no. 1, pp. 400-405.

Santos, JP, Hontanon, E, Ramiro, E & Alonso, M 2009, 'Performance evaluation of a high-resolution parallel-plate differential mobility analyzer', *Atmospheric Chemistry and Physics*, vol. 9, no. 7, pp. 2419-2429.

Savitzky, A & Golay, MJE 1964, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', *Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639.

Savorani, F, Tomasi, G & Engelsen, SB 2010, 'icoshift: A versatile tool for the rapid alignment of 1D NMR spectra', *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190-202.

Scott, SM, James, D & Ali, Z 2006, 'Data analysis for electronic nose systems', *Microchimica Acta*, vol. 156, no. 3-4, pp. 183-207.

Sferopoulos, R 2009, 'A Review of Chemical Warfare Agent (CWA) Detector Technologies and Commercial-Off-The-Shelf Items', *Defence Science and Technology Organisation*, vol. Department of Defence, Australian Government.

Singh, S & Singh, M 2003, 'Explosives detection systems (EDS) for aviation security', *Signal Processing*, vol. 83, pp. 31-55.

Skov, T, van den Berg, F, Tomasi, G & Bro, R 2006, 'Automated alignment of chromatographic data', *Journal of Chemometrics*, vol. 20, no. 11-12, pp. 484-497.

Tauler, R 1995, 'Multivariate curve resolution applied to second order data', *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 133-146.

Tauler, R, Kowalski, B & Fleming, S 1993, 'Multivariate curve resolution applied to spectral data from multiple runs of an industrial process', *Analytical Chemistry*, vol. 65, no. 15, pp. 2040-2047.

Tibshirani, R 1996, 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267-288.

Tomasi, G, van den Berg, F & Andersson, C 2004, 'Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data', *Journal of Chemometrics*, vol. 18, no. 5, pp. 231-241.

Vergassola, M, Villermaux, E & Shraiman, BI 2007, 'Infotaxis as a strategy for searching without gradients', *Nature*, vol. 445, no. 7126, pp. 406-409.

Windig, W, Gallagher, NB, Shaver, JM & Wise, BM 2005, 'A new approach for interactive self-modeling mixture analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, pp. 85-96.

Wold, S, Esbensen, K & Geladi, P 1987, 'Principal Component Analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52.

Wold, S, Kettaneh-Wold, N & Skagerberg, B 1989, 'Nonlinear PLS modeling', *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 53-65.

Wold, S, Sjöström, M & Eriksson, L 2001, 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 109-130.

Yee, E 2008, 'Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference', *Boundary-Layer Meteorology*, vol. 127, pp. 359-394.

Yee, E 2009, 'Probability law of concentration in plumes dispersing in an urban area', *Environmental Fluid Mechanics*, vol. 9, pp. 389-407.

Yee, E & Biltoft, CA 2004, 'Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles', *Boundary-Layer Meteorology*, vol. 111, pp. 363-415.

Yee, E & Chan, R 1997, 'A simple model for the probability density function of concentration fluctuations in atmospheric plumes', *Atmospheric Environment*, vol. 31, pp. 991-1002.

Yee, E, Wang, B-C & Lien, F-S 2009, 'Probabilistic model for concentration fluctuations in compact-source plumes in an urbant environment', *Boundary-Layer Meteorology*, vol. 130, pp. 169-208.

Zamora, D & Blanco, M 2012, 'Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration', *Analytica Chimica Acta*, vol. 726, no. 0, pp. 50-56.

Zarzhitsky, D, Spears, DF & Spears, WM 2005, 'Swarms for chemical plume tracing', in *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, pp. 249-256.

Zhang, ZM, Chen, S & Liang, YZ 2010, 'Baseline correction using adaptive iteratively reweighted penalized least squares', *Analyst*, vol. 135, no. 5, pp. 1138-1146.

Zheng, P, Harrington, PdB & Davis, DM 1996, 'Quantitative analysis of volatile organic compounds using ion mobility spectrometry and cascade correlation neural networks', *Chemometrics and Intelligent Laboratory Systems*, vol. 33, pp. 121-132.

Chapter 9

List of publications and conferences

9.1 Publications

9.1.1 Journals

- **Pomareda, V**, Lopez-Vidal, S, Calvo, D, Pardo, A & Marco, S 2013, 'A novel differential mobility analyzer as a VOC detector and multivariate techniques for identification and quantification', *Analyst*, no. 138, pp. 3512-3521.

- **Pomareda, V**, Guamán, AV, Mohammadnejad, M, Calvo, D, Pardo, A & Marco, S 2012, 'Multivariate curve resolution of nonlinear ion mobility spectra followed by multivariate nonlinear calibration for quantitative prediction', *Chemometrics and Intelligent Laboratory Systems*, vol. 118, no. 0, pp. 219-229.

- **Pomareda, V**, Calvo, D, Pardo, A & Marco, S 2010, 'Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra', *Chemometrics and Intelligent Laboratory Systems*, vol. 104, no. 2, pp. 318-332.

9.1.2 Proceedings and workshops

- Hernandez-Bennets, V, Lilienthal, AJ, Khaliq, AA, **Pomareda, V** & Trincavelli, M 2013, 'Towards Real-World Gas Distribution Mapping and Leak Localization Using a Mobile Robot with 3D and Remote Gas Sensing Capabilities', in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe (Germany), p. to appear **(Best Service Robotics Paper Award)**.

- Hernández-Bennetts, V, Lilienthal, AJ, Khaliq, AA, **Pomareda, V** & Trincavelli, M 2012, 'Gasbot: A mobile robotic platform for methane leak detection and emission monitoring', in *IROS Workshop on Robotics for Environmental Monitoring*, Vilamoura, Algarve, Portugal, p. to appear.

- **Pomareda, V** & Marco, S 2011, 'Chemical plume source localization with multiple mobile sensors using Bayesian inference under background signals', in *AIP Conference Proceedings 14th International Symposium on Olfaction and Electronic Nose*, New York City, New York, USA, pp. 149-150.

- Marco, S, **Pomareda, V**, Pardo, A, Kessler, M, Goebel, J & Mueller, G 2009, 'Blind Source Separation For Ion Mobility Spectra', in *Olfaction and Electronic Nose, Proceedings*, vol. 1137, eds M Pardo & G Sberveglieri, Amer Inst Physics, Melville, pp. 551-553.

- Montoliu, I, **Pomareda, V**, Kalms, A, Pardo, A, Gobel, J, Kessler, M, Muller, G & Marco, S 2009, 'Resolution of Ion Mobility Spectra for the Detection of Hazardous Substances in Real Sampling Conditions', in *Olfaction and Electronic Nose, Proceedings*, vol. 1137, eds M Pardo & G Sberveglieri, Amer Inst Physics, Melville, pp. 576-578.

9.2 Participation in conferences

9.2.1 Oral

- **Pomareda, V**, Hernández-Bennetts, V, Khaliq, AA, Trincavelli, M, Lilienthal, AJ & Marco, S 2013, 'Chemical source localization in real environments integrating chemical concentrations in a probabilistic plume mapping approach', in *15th International Symposium on Olfaction and Electronic Nose (ISOEN 2013)*, Daegu, Korea, July 2013.

- **Pomareda, V** & Marco, S 2011, 'Chemical plume source localization with multiple mobile sensors using Bayesian inference under background signals', in *14th International Symposium on Olfaction and Electronic Nose (ISOEN 2011)*, New York City, New York, USA, May 2011.

- **Pomareda, V**, Calvo, D, Pardo, A, Rodríguez, J, Montoya, E & Marco, S 2009, 'Blind Source Separation for Ion Mobility Spectrometry: Application to a Luggage Scanner', in *The 18th Annual Conference on Ion Mobility Spectrometry (ISIMS 2009)*, Thun, Switzerland, July 2009.

9.2.2 Poster

- **Pomareda, V**, Calvo, D, Pardo, A & Marco, S 2010, 'Hard-Modeling in Non-negative Matrix Factorization: Application to Ion Mobility Spectra', in *12th Conference on Chemometrics and Analytical Chemistry (CAC 2010)*, Antwerp, Belgium, October 2010 **(poster awarded)**.

- Marco, S, **Pomareda, V**, Pardo, A, Kessler, M, Goebel, J & Mueller, G 2009, 'Blind Source Separation For Ion Mobility Spectra', in *13th International Symposium on Olfaction and Electronic Nose (ISOEN 2009)*, Brescia, Italy, July 2009.

- Montoliu, I, **Pomareda, V**, Kalms, A, Pardo, A, Gobel, J, Kessler, M, Muller, G & Marco, S 2009, 'Resolution of Ion Mobility Spectra for the Detection of Hazardous Substances in Real Sampling Conditions', in *13th International Symposium on Olfaction and Electronic Nose (ISOEN 2009)*, Brescia, Italy, July 2009.