Signal Processing for Artificial Olfaction

Pierre Maho

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **SIGNAL IMAGE PAROLE TELECOMS**

Arrêté ministériel : 25 mai 2016

Présentée par

Pierre MAHO

Thèse dirigée par **Pierre COMON,** Université Grenoble Alpes et codirigée par **Simon BARTHELMÉ,** CNRS

préparée au sein du **Laboratoire Grenoble Images Parole Signal Automatique**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)**

Traitement du Signal pour l'Olfaction Artificielle
Signal Processing for Artificial Olfaction

Thèse soutenue publiquement le **29 juin 2020**,
devant le jury composé de :

Monsieur Jérôme IDIER
Directeur de recherche, CNRS Délégation Bretagne Pays de Loire, Examinateur
Monsieur Said MOUSSAOUI
Professeur des Universités, Ecole Centrale Nantes, Rapporteur
Monsieur David BRIE
Professeur des Universités, Université de Lorraine, Rapporteur
Madame Yanxia HOU-BROUTIN
Chargée de recherche, CNRS Délégation Alpes, Examinatrice
Monsieur Rasmus BRO
Professeur, Université de Copenhague (Danemark), Examinateur
Monsieur Thierry LIVACHE
Directeur scientifique, Aryballe, Examinateur
Monsieur Nicolas LE BIHAN
Directeur de recherche, CNRS Délégation Alpes, Président
Monsieur Bertrand BOURLON
CEA LETI, Invité

Signal Processing for Artificial Olfaction

Pierre Maho

September 28, 2020

Please do not circulate this thesis.

This document will be available online only after:

July 1, 2021

Acknowledgement

Dear reader, after 3,5 years of work, here are the fruits of my labour; ~200 pages about artificial olfaction and more modestly about an electronic nose. 3,5 years of my life, what a relief! But I could not have made it without all the people who supported me during this long term endeavour. During the next few lines, I will switch to French to acknowledge them.

Commençons par ceux sans qui cette thèse n'aurait pas été possible, mes encadrants Pierre Comon et Simon Barthelmé. Merci à tous les deux de m'avoir épaulé et fait confiance pendant ces 3 ans. J'ai toujours pu compter sur vous et votre porte était toujours grande ouverte. Je mesure ma chance et tacherais de faire bon usage de ce que vous m'avez transmis. Un grand merci aussi à mes encadrants officieux côté Aryballe, Thierry Livache et Cyril Herrier. Un traiteur de signaux ne va jamais bien loin sans les experts du domaine dont proviennent les fameux signaux. Vous nous avez donc apporté le regard de chimiste qui nous faisait tant défaut. Merci à tous les 4 d'avoir cru à ce projet!

Un grand merci aux membres du jury, à ceux qui ont pu se déplacer et à ceux qui étaient par visio. Coronavirus oblige, la soutenance était quelque peu spéciale et plutôt inédite dans son déroulement. Merci à David Brie et Said Moussaoui, dans leur rôle de rapporteur. Merci à Yanxia Hou-Broutin, Thierry Livache, Jérôme Idier et Rasmus Bro (thanks!), dans leur rôle d'examinateur. Et enfin, merci à Bertrand Bourlon d'avoir accepté notre invitation et à Nicolas Le Bihan d'avoir tenu le rôle de président.

Une thèse, c'est long et court à la fois...

... Tout commence par un bureau, dans lequel vont et viennent les grands de ce monde. Et dans le D1110, du beau monde est passé! Marc l'athlète intrépide. Alex M le communicant. Cosme, un ch'ti biloute en devenir qui a de l'avenir. Lorenzo, l'italien au franc-parler. Reda, débatteur invétéré.

... Puis, il arrive que l'on sorte de son bureau, qu'on déambule dans les couloirs, qu'on erre dans les cafet'. Ainsi, d'autres amitiés naissent et se tissent au fil des années. Marion, Fabrice & Miya, vous pourrez toujours compter sur moi pour ne pas devoir finir les restes. Imane, la boxeuse-grimpeuse-artiste-docteure-patissière-(bientôt)barwoman-... Kevin, le grand pont entre la City et les Zadistes. Raph, mon éternel maître de stage et aventurier parti au bout du monde chasser le manchot. Jeanne & Louis, au plaisir de bivouaquer ensemble maintenant que je suis bien équipé grâce à vous, Jeanne tes visites et nos discussions dans mon bureau

me manquent déjà. ManuE, quand bébé doctorant devient grand. Pedro Luis, ma version brésilienne améliorée. Malik, l'enjoliveur d'histoires. Roza, la relève au Gipsa est assurée. Fanny, notre ascension du Moucherotte restera dans nos mémoires, puisse t-on le regravir ensemble un jour! Et tant d'autres que je ne peux que citer par manque de place: Julien, Aziliz, Pierre, Valentin, Maxime, Fateme, Elaheh, Ludo, Violette, Florent, Miguel, Marielle, Taïa, Dawood, Maria, Firas, Hélène, Sonia, Anne...

... Il n'est pas rare aussi qu'on ose mettre le pied un peu plus loin, voire même dehors! Merci à la team Foot, pour tous ces bons moments passés à s'entraîner, à gagner et à perdre, souvent. Merci de m'avoir fait confiance pour être capitaine de la fameuse GIPSA2, ce fut un honneur. Merci à Sylvain & Silvain, à Xavier, Hazem, Andrei, Jérémy, Raul, Mika, Rattena, Aurélien, Alex H, Rémy, Haris, Cyprien, Cosme,... Merci à l'équipe plateforme du GIPSA, pour votre bon esprit et votre bonne humeur dont vous seuls avez le secret au labo. Merci à Sophie, Rémy, Aurélien, Anton, Sylvain & Silvain. Merci à la team Aryballe, que je rejoins et à qui je dois aussi la réussite de cette thèse. Merci à Johanna, Etienne, Guillaume, Romain, Krishna,... Et à la team CREAB, Jonathan, Charlotte, Yanxia et Arnaud.

... Enfin, même si la thèse occupe l'esprit, elle n'occupe pas exclusivement votre vie. Merci à mes parents et à mes frères, les retours dans le Nord ont toujours une saveur et une nostalgie particulières grâce à vous. Merci à mon père et à mon frère Thomas d'avoir fait la route jusque Grenoble malgré les incertitudes qui auront duré jusqu'au dernier moment à cause du Covid. Merci à Marylène, que j'attends sur les pistes et dont l'amitié n'a pas souffert de la distance. Merci à Anne So, pour tout ce qu'on a pu vivre ensemble, pour tes nombreuses visites à Grenoble, pour ne pas m'oublier malgré la distance et pour m'avoir fait confiance en me laissant Loustic. Merci à Loustic, un être qui m'est cher et qui m'accompagne au quotidien depuis 6 ans, à qui je dois au moins 2500 heures de plein air qui m'auront permis de m'évader un peu, et de rencontrer aussi des personnes qui comptent désormais. Je pense notamment à Aurélie, voisine et désormais amie à qui je peux dire un grand merci de m'avoir fait rencontrer Yanicka. Enfin, merci à Yanicka d'avoir accepté de me suivre dans cette aventure qu'est la fin de thèse. Pas facile de s'engager avec quelqu'un dont l'esprit est bien occupé, mais tu l'as fait et je ne t'en remercierai jamais assez. Merci pour ton soutien indéfectible et pour notre complicité au quotidien. Longue vie à nous!

Finally, a big thanks to Patrick Süskind for the starts of my chapters.

Contents

List of notations

$\boldsymbol{A}$	Matrix
$\boldsymbol{A}^{-1}$	The inverse of the matrix $\boldsymbol{A}$
$\boldsymbol{A}^{\mathrm{T}}$	Transposed matrix
$\boldsymbol{x}$	Column vector
$\boldsymbol{x}^{\mathrm{T}}$	Row vector
$\boldsymbol{a}_j$	j^{th} column of $\boldsymbol{A}$
$\boldsymbol{a}_i^{\mathrm{T}}$	i^{th} row of $\boldsymbol{A}$
a_{ij}	i^{th} row and j^{th} column of $\boldsymbol{A}$
$\hat{\boldsymbol{A}}$	Estimate
$\boldsymbol{I}$	Identity matrix
$\boldsymbol{A} \geqslant 0$	Matrix with non-negative terms
$\mathrm{rank}(\boldsymbol{A})$	Rank of the matrix $\boldsymbol{A}$
$\mathrm{diag}(\boldsymbol{x})$	Diagonal matrix formed by $\boldsymbol{x}$
$\boldsymbol{A} \boxdot \boldsymbol{B}$	Hadamard (entrywise) product
$\boldsymbol{A} \boxslash \boldsymbol{B}$	Hadamard (entrywise) division
$\|\boldsymbol{x}\|_1$	ℓ_1-norm of $\boldsymbol{x}$
$\|\boldsymbol{x}\|_2$	ℓ_2-norm or Euclidean norm of $\boldsymbol{x}$
$\|\boldsymbol{A}\|_F$	Frobenius norm of $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{\max}$	Max-norm of $\boldsymbol{A}$
r	VOC index
p	Chemical sensor index
n	Experiment or sample index
t	Time index
R	Number of VOCs
P	Number of chemical sensors
M	Number of sensing materials
N	Number of experiments or samples
c	Concentration
a	Adsorption constant
d	Desorption constant
k	Affinity constant

A	Molecule
S	SPRi sensitivity (variation of reflectivity induced by a variation of refractive index)
$\frac{\partial n}{\partial c}$	change in refractive index of the VOC with change in concentration
m	Molecular mass of the VOC
L_z	Penetration depth of the evanescent wave
e_b	Thickness of the medium formed by the adsorption of the VOC on the sensing surface
θ	Fraction of occupied sites
θ^{eq}	Fraction of occupied sites at equilibrium
τ^a	Adsorption constant in the simplified model
τ^d	Desorption constant in the simplified model
t_s	Beginning of the VOC injection
t_e	End of the VOC injection
[A]	Concentration of the molecule A
$[\text{A}]_s$	Surfacic concentration of the molecule A
ΔR	Delta-reflectivity
y	Response of the instrument
ℓ	Class (or label)
$\boldsymbol{L}$	Class-membership matrix (each row contains a single 1 indicating the class membership)
$\boldsymbol{X}$	Signature matrix
$\boldsymbol{X}_{\text{cal}}$	Signature matrix of a calibrant
$\mathring{\boldsymbol{X}}$	Signature matrix in sensing materials space
$\boldsymbol{C}$	Concentration matrix
$\boldsymbol{K}$	Affinity matrix
$\boldsymbol{S}$	Covariance matrix
$\boldsymbol{\mu}$	Mean
$\boldsymbol{S}_w$	Within-class covariance matrix
$\boldsymbol{S}_b$	Between-class covariance matrix
$\mathcal{S}$	Support (position of non-null coefficients)
σ_n	Standard deviation of noise
$\boldsymbol{d}$	Drift direction
$\boldsymbol{\alpha}$	Weight vector (weighting the drift for each class)
$\boldsymbol{\pi}$	Probability vector
$\mathbb{E}(\boldsymbol{x})$	Expectation of $\boldsymbol{x}$

List of acronyms

VOC	Volatile Organic Compound
GC	Gas Chromatography
MS	Mass Spectrometer
PID	PhotoIonization Detector
FID	Flame Ionization Detector
MOS	Metal Oxide Semiconductor
TGS	Taguchi Gas Sensor
MIP	Molecularly Imprinted Polymer
SPRi	Surface Plasmon Resonance imaging
TM	Transverse Magnetic
TE	Transverse Electric
DNA	DeoxyriboNucleic Acid
PEEK	PolyEtherEtherKetone
ALS	Alternating Least Squares
SVM	Support Vector Machine
k-NN	k Nearest Neighbors
SOM	Self-Organized Maps
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
PLS	Partial Least Squares
CC	Component Correction
CPCA	Common Principal Component Analysis
OSC	Orthogonal Signal Correction
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
MR	Moving Recentering
EM	Expectation-Maximization
EMC^2	Expectation-Maximization Component Correction
BFGS	Broyden-Fletcher-Goldfarb-Shanno
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
CV	Cross-Validation
SSE	Sum of Squared Errors
SNR	Signal-to-Noise Ratio

CS	Calibrant Scenario
MSS	Multi-Session Scenario
BS	Blind Scenario
EFOS	Equal Frequency of Occurrence Scenario
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives
INAC	Institut NAnosciences et Cryogénie
SyMMES	Systèmes Moléculaires et nanoMatériaux pour l'Energie et la Santé
CREAB	Chimie pour la Reconnaissance et l'Etude des Assemblages Biologiques
CNRS	Centre National de la Recherche Scientifique
GIPSA-Lab	Grenoble Images Paroles Signal Automatique laboratoire
FUI	Fond Unique Interministériel
WISE	Wellness and medical dIagnostics olfactory SEnsors
BPI	Banque Publique d'Investissement

Structure of the thesis

In this thesis, we study a class of instruments broadly inspired from natural Olfaction, namely electronic noses. While our nose helps us to identify and describe an odorant, an electronic nose aims at analyzing any volatile compound in the environment. Their main feature is the use of several non-specific chemical sensors which allow them to be sensitive to a broad range of different compounds. However, this non-specificity is also behind various Signal Processing challenges. For instance, the instrument is always used in combination with Machine Learning algorithms to differentiate between two compounds A and B. This is due to the fact that all the non-specific sensors will likely respond positively to A or B so, classifying a future sample is simply not a human task. In addition, in the presence of a gas mixture, the response of a chemical sensor is influenced by all the compounds composing the mixture. The measurement is affected in a non-trivial way by the individual response of each component and by their concentration. Designing Source Separation methods can then help to identify and quantify each compound from the measurement. Finally, chemical sensors are also prone to drift over time, meaning that their response towards a compound A can change from a day d to a day $d+1$. In this case, the development of correction methods is compulsory to reuse what is learnt during day d. All these challenges are addressed in this thesis with an optoelectronic developed by Aryballe (France, Grenoble).

In Chapter 1, we first introduce the traditional techniques for analyzing volatile compounds. Then, we define the concept of electronic noses by listing in particular the main challenges and some of the existing technologies. We provide a short critical review of the literature by pointing out some negative aspects. Afterwards, we introduce the optoelectronic nose used in this thesis and provided by Aryballe. Some of the results already obtained with this instrument are reviewed and we finally conclude by describing the different issues studied in this thesis.

Chapter 2 is a small chapter gathering all the setups used. We introduce an automatic valve and two robotic platforms designed and developed during the thesis.

In Chapter 3, we study how to relate the response of the instrument to some physico-chemical parameters, such as concentration. We first describe a model of the response over time when a pure compound is measured. In a second part, the model is validated using two different real data sets. The parameters of the model are also evaluated as potential features for

classification purpose. In the last part, we formulate a model of the response at equilibrium when a gas mixture is measured. Some of the theoretical properties of this model are studied. We also develop an algorithm in a Blind Source Separation framework and we validate it on simulated data.

In Chapter 4, one of the main problems of electronic noses is studied, namely drift issue. We first introduce the main causes and consequences of the drift. Then, we describe some correction methods used in the literature by classifying them according to different scenarios. Afterwards, we develop two correction methods which are based on a model. The main feature of these two methods is that they do not require any label from the new measurement sessions. We evaluate these methods on a real data set acquired over 9 months and affected by drift. All the introduced methods are then compared and we show that the two proposed algorithms perform fairly and compensate, at least partly, for the drift.

In Chapter 5, we show the stereoselectivity of the instrument for two pairs of enantiomers (Carvone and Limonene). We first introduce the problem of differentiating two enantiomers in analytical chemistry and the different existing methods to do that. Then, we propose an experimental methodology to reliably evaluate the stereoselectivity of an electronic nose. This methodology is applied to Aryballe's eNose and we highlight some potential confounds which could refute comparable studies. We finally deeply investigate the origin of the stereoselectivity and we successfully isolate the chemical sensors responsible for the discrimination.

In Chapter 6, we get the instrument out the lab by placing it on a robot. We first introduce the robot olfaction field and its main challenges. Two of them are explored in this chapter: classification and Source Separation. To that end, we design a real-time pipeline based on a sparse linear model capable of both classifying and unmixing signals. We validate the proposed algorithm on different data sets generated during the thesis: classification of 12 volatile compounds disseminated in the environment, unmixing of binary and ternary mixtures, etc...

Chapter 7 summarizes all the results obtained during the thesis and discusses some future prospects.

1 Introduction

Contents

"For people could close their eyes to greatness, to horrors, to beauty, and their ears to melodies or deceiving words. But they could not escape scent. For scent was a brother of breath. Together with breath it entered human beings, who could not defend themselves against it, not if they wanted to live. And scent entered into their very core, went directly to their hearts and decided for good and all between affection and contempt, disgust and lust, love and hate. He who ruled scent ruled the hearts of men."

Patrick Süskind, *Perfume, the Story of a Murderer*[1].

In his book, Patrick Süskind describes a man having a supernatural sense of smell. Jean Baptiste Grenouille had the "ability to see right through paper, cloth, wood, and even through brick walls and locked doors. Without ever entering the dormitory, he knew how many of her wards - and which ones - were in there". Unfortunately for him, this natural gift turned rapidly into an obsession and this obsession into a crime when he decided to kill a woman (and many others after her). This terrible act was motivated by an idea: capturing the fragrance escaping from her body. This idea was motivated by a desire and this desire by an odor and this odor by an odorant. But what is the pathway between the sensation and the initial odorant?

Let us start at the beginning. An odorant is simply a set of molecules which can be evaporated in the air (or which are soluble in water for aquatic animals). These molecules must be light enough to evaporate but can be really different in shape, size and structure. Through respiration, these molecules are inhaled by the nose and meet the olfactory epithelium, a small area (1-5 cm^2) on the ceiling of the nasal cavity. When meeting the epithelium, the molecules meet several millions of Olfactory Receptor Neurons (ORNs) and bind to some of them. Each ORN, out of ~400 different types in humans, shows a cross-sensitivity towards the molecules meaning that it can interact with several molecules and a single molecule can interact with several ORNs [Mal+99]. If a molecule can bind to the ORN, then the neuron is activated and converts this interaction into an electrical signal which is transmitted to the brain. These messages are received by the olfactory bulb and the combination of activated neurons is decoded by the brain which in turn translates the odorant into a known odor [Mal+99]. In fact, regarding the decoding process, a lot of mysteries are still to be unraveled. For instance, how many odorants can the brain discriminate? What is the relation between the odorant and its odor? Is it only related to the physico-chemical properties of the odorant? In this case, why can two enantiomers smell completely different (two enantiomers are almost identical molecules, but we will come back to this notion in Chapter 5)? These questions are still largely debated by researchers and there exists actually no theory giving full satisfaction (see for instance the debate by interposed PNAS letters between Turin's group and Block's group [Tur+15; Blo+15a], after the article of Block et al.[Blo+15b] about the vibrational theory of olfaction). This thesis will... not try to answer them, not even a little bit.

[1]Translated from the German by John E. Woods.

No, this thesis is not about natural olfaction but artificial olfaction. Nature did wonders to allow humans to be sensitive to surrounding odorants but how do humans to measure, identify and quantify an odorant by instruments?

1.1 Traditional techniques for VOC analysis

The term "odorant" is in fact a bit misleading for artificial olfaction since it could be unintentionally related to the perception ("odor"). In the following, we prefer to use the term **VOC**, for **Volatile Organic Compound** (a commonly used term), for describing a molecule present in an odorant. Technically, a VOC is a compound containing carbon which easily evaporates. Since an odorant is composed of different VOCs, we refer to odorant as **gas mixture** (it is also sometimes called a multicomponent gas). Odorant and odor will be kept only in the case where there is no ambiguity (as in the next section).

1.1.1 Dog's snout and human nose as tools

There is still no instrument capable of what a human nose can do (and even less what a dog's snout can do). Therefore, human nose is still a widely used technique in food and environmental industries. In practice, it consists in using sensory panelists who fill out some questionaries about the quality of food, perfumes, cosmetics or chemical products [CLH12]. Since we do not know actually how to relate the odorant[2] to the odor, the use of human panels is still crucial regarding for instance the hedonic (like/dislike) property of an odorant.

Dogs show a larger diversity of olfactory receptors (~1000, [AY05]) and a much larger surface area of olfactory epithelium (depending on the dog breed, [Qui+12]) than humans. They have been used in a broad variety of scent-related tasks for hundreds of years. Nowadays, they are routinely used to detect explosives or drugs but are also deployed in the field to search for victims in disaster areas [GT03]. Recently, several papers have drawn attention to the use of dogs in human medicine (for instance, cancer or diabetes detection) [MM10]. Due to their incredible sense of smell and to their capacity to be trained, dogs have still a bright future.

These techniques are very useful, and will probably still be for a long time, since technology is still far from natural systems (*e.g.* ability to work in everyday life conditions, fast measurement, link between odorant and feelings, etc...). However, both humans and dogs, can suffer from fatigue and subjectivity, and be affected by physical and mental conditions [Pea+06]. For instance, the demonstration of dogs' ability, especially for medical purposes, are not free from experimental confounds (*e.g.* dogs can detect tiny changes of human body language and can then interpret them as clues) [JHFT13; JHFT17]. For humans, the odor of an odorant suffers from inter-individual variations and depends on culture and gender [Fer+13]. Besides, both humans and dogs require intensive and expensive training and cannot tell exactly the amount of an odorant.

[2] In this subsection, we kept "odorant" and "odor" since the subsection directly relates to natural olfaction.

To avoid the problems of natural systems and to provide a more objective measurement of an odorant, analytical instruments have been developed. In the next section, we detail the gold standard: Gas Chromatography - Mass Spectrometer.

1.1.2 Gas Chromatography - Mass Spectrometer

Gas Chromatography (GC) is the gold standard for separating a gas mixture. The main principle of GC relies on the use of a column (a tube). The most used type of columns is the capillary column. It is a long tube (30-60 m, so it is a rolled-up tube) in stainless steel whose internal surface is covered with a molecule, called the stationary phase. The gas mixture is injected in this column through a carrying gas (usually Helium), called the mobile phase. All the VOCs composing the gas mixture will have different affinities with the mobile and stationary phases. For instance, assuming that the gas mixture is composed of two VOCs, say A and B, if A has more affinity with the stationary phase than B, then A will spend more time in the column than B (interacting takes time). So B will leave the column first, followed by A. We say that the retention time of B is lower than the retention time of A. More details about GC can be found in [MMS19]. A simplified representation of the principle of GC is reported in Figure 1.1.

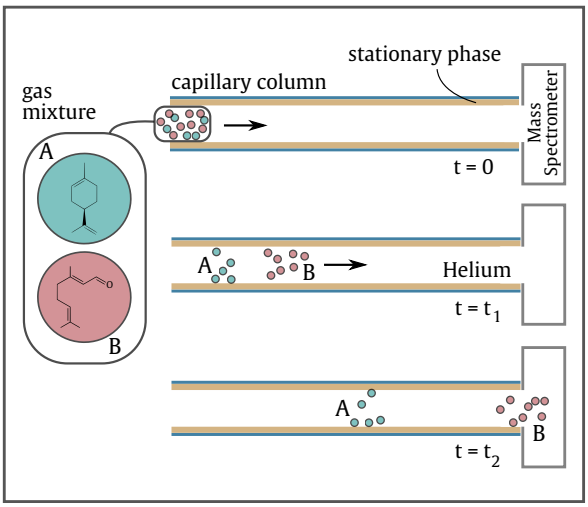


Fig. 1.1.: Principle of Gas Chromatography. A gas mixture composed of A and B is injected through a carrying gas (mobile phase) in a capillary column whose internal surface is covered with a molecule (stationary phase). A and B will interact differently with the mobile and stationary phases. Here, B interacts less so it leaves first the column.

In the output of a GC, the components of the gas mixture have been physically separated using their retention time. However, they are not yet detected, or identified, or quantified. In other words, we do not know when they left the column, which VOC and which amount of VOC left the column. For that, we need another instrument and there exist several different devices, such as Flame Ionization Detector or Mass Spectrometer.

The principle of GC is simple compared to **Mass Spectrometer** (MS) which is a bit more technical. MS is a destructive technique which performs like a scale: it identifies molecules using their mass. Its principle relies on the separation of electrically charged molecules as a function of the ratio $\frac{m}{z}$ (m the mass, z the charge). MS is composed of 3 steps: ionization, mass separation, detection. Each one of these steps can be performed by several different instruments. Detailing each instrument can make things more complicated whereas the principle remains the same. So we detail only one way for performing each one of the 3 steps (see [DA06] for a comparison of methods). The first step is ionization. During ionization, the molecule, say B since it left first the GC, is bombarded with free electrons (this technique

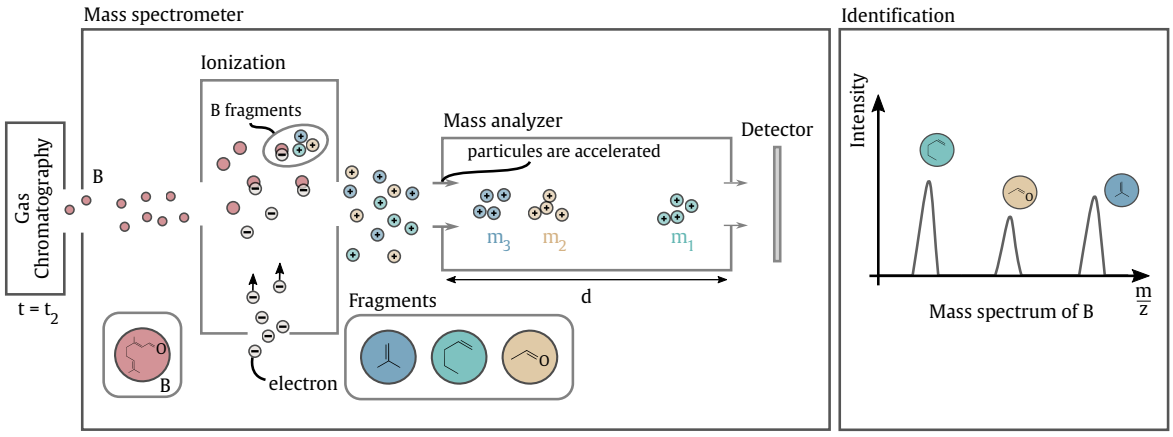


Fig. 1.2.: Principle of Mass spectrometer, using Electron Ionization and Time Of Flight. The compound B is fragmented by being bombarded with free electrons. The fragments are then sorted by their mass by being accelerated in a field-free environment. The lighter the fragment is, the lower the time to travel a certain distance is. Therefore, the lightest fragment hits the detector first, followed by the second lightest, etc... At the end, we obtain the mass spectrum of B, characteristic of the compound.

is called Electron Ionization) causing the fragmentation of B into charged molecules (ions). The way B fragments into smaller molecules is reproducible and characteristic of B (provided that the experimental parameters remain the same). Since the resulting pieces depend on B, a way to obtain a pattern representing B is to sort the pieces by their mass: it is the mass spectrum of B.

Let us illustrate the mass spectrum by means of analogy, an analogy similar to the one proposed by [Mab88]. Assume that we want to discriminate two toys that we cannot see, say a doll and a toy car. Assume that we can break each toy in a reproducible manner: a doll will break into pieces such as arms, legs and head, and a toy car into tires, engine and doors. Then, we weigh each piece and draw a chart: in x-axis the weight in grams and in y-axis the number of pieces having this weight. For instance, the toy car will have a 4 high peak at 100g corresponding to the 4 tires and the doll will have a 2 high peak at 50g corresponding to the arms. At the end, we obtain the mass spectra of the doll and the toy car. They will look different and can be compared to a data base containing a lot of spectra of toys. Mass spectrometer will do the same but at a molecular level.

A way to sort the fragments of B by their mass is to use the Time Of Flight. Time Of Flight relies on the principle that any particule accelerated in a field-free environment (*i.e.* no friction) will take a time t to travel a distance d which is directly proportional to the ratio $\frac{m}{z}$: the lighter the molecule is, the lower t is. So, by applying a voltage to the pieces of the fragmented molecule B, each piece is accelerated towards a detector (detection step) which counts the number of ions hitting it (generally through a current intensity). Each piece will take some time to reach the detector, so we have a spectrum over time. Since the time of flight is directly proportional to $\frac{m}{z}$, the spectrum over time is simply converted into the mass spectrum (by knowing the distance d and the voltage).

By comparing this spectrum with a large data base of spectra, we identify the VOC as B. One of the advantages of MS and one of the reasons to use it with GC is clearly its time of acquisition, acquiring a mass spectrum requires only ∼200 ms. A simplified representation of MS, using Electron Ionization and Time Of Flight, is reported in Figure 1.2.

The coupling of GC and MS is called GC-MS. The two instruments are quite complementary: GC allows to separate the components of a gas mixture over time while MS identifies and quantifies each VOC leaving at any time GC.

The result of a GC-MS is called a chromatogram: it represents the amount of molecules measured by the detector over time and looks generally as a series of peaks (each VOC leaving the column one-by-one). In fact, each point corresponds to a mass spectrum which has been summed. The information about the identity of the VOC at time t is contained in this spectrum while the integration of the peak tells us about the concentration of the VOC in the gas mixture. A representation of a chromatogram is reported in Figure 1.3.

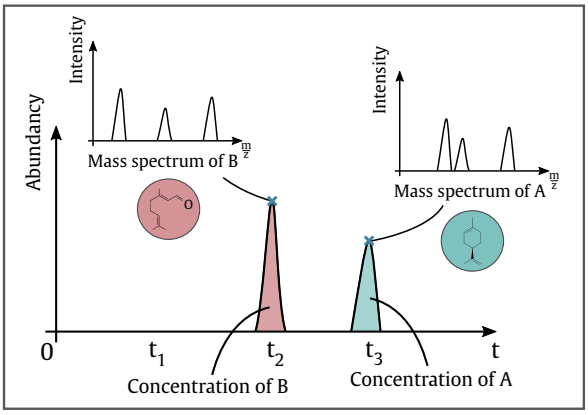


Fig. 1.3.: Result of a GC-MS: a chromatogram. It represents the amount of molecules measured by the detector over time. Each time point is a mass spectrum, enabling the VOC identification.

GC-MS is the gold standard for analyzing precisely and reliably a gas mixture with high-sensitivity and high-selectivity. In other words, GC-MS is able to detect small changes of VOC concentration (high-sensitivity) and can easily detect a given VOC in presence of interferent VOCs (high-selectivity). However, it suffers from several drawbacks. GC-MS is clearly not portable, is quite expensive, works in quite controlled conditions and requires qualified personnel to be used. All these reasons restrict the use of these instruments to laboratories in analytical chemistry or large companies.

1.1.3 Gas Chromatography - Olfactometry

Another method is also regularly used in odorant analysis: Gas Chromatography - Olfactometry (GC-O). It combines the separation power of GC with human perception. Instead of using a detector such as MS, a human nose inhales the components of the odorant leaving the GC. It generally consists in replying to questions such as detection of an odor, duration of odor activity, quality or intensity of the perceived odor [DED06]. The idea is then to correlate sensory responses with volatile chemicals. For instance, it helps to understand which molecules are responsible for unpleasant odors in a complex odorant. Th result of a GC-O is reported in Figure 1.4.

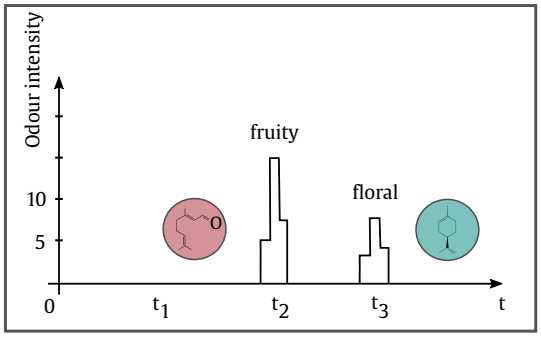


Fig. 1.4.: Example of a result from GC-O. A human nose inhales the components of the odorant leaving the GC, giving information about the separated VOCs (intensity, perceived odor, etc...).

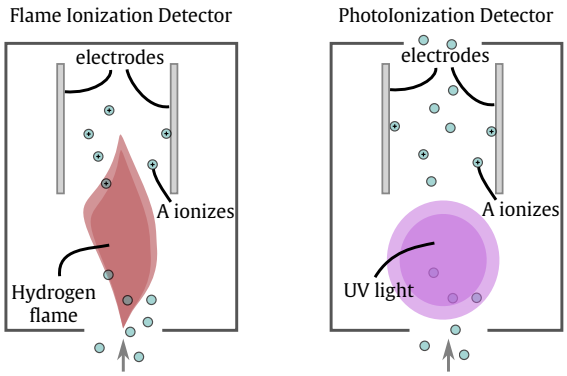


Fig. 1.5.: Principle of Flame Ionization Detector (FID) and PhotoIonization Detector (PID). The compound is ionized using either a Hydrogen flame (FID) or UV light (PID). The ions are captured by the electrodes and the generated current is proportional to VOC concentration.

1.1.4 Ionization-based detectors

There are other types of detectors based on ionization: **Flame Ionization Detector** (FID) [MWD58] and **PhotoIonization Detector** (PID) [RAE14]. For FID, the VOC is ionized by using a flame generated by the combustion of a Hydrogen flow. Then, the formed ions go to two electrodes and generate a current. This current is proportional to VOC concentration. FID is a destructive method, so the VOC cannot be reused for further analyses. For PID, the principle remains the same but the ionization is performed using a light source with high-energy photons (generally UV light), which allows to reuse partly the VOC. Both techniques are not selective so they cannot identify the VOC but they can quantify almost any organic compound. For this reason, FID and PID are often used as a complementary instrument to GC.

1.2 Electronic noses

The tools described above for analyzing a VOC suffer from some drawbacks. Natural systems are expensive to train and employ, lack objectivity and are prone to fatigue. GC-MS is quite expensive, cannot be used on a large scale or for field monitoring and requires qualified personnel. FID and PID do not allow the identification of the VOC. In comparison, another kind of tools is particularly appealing for its portability, low cost, easy-to-use and its ability to identify a broad variety of VOCs. This class of instruments takes inspiration from the mammalian olfactory system and is usually referred to as **electronic nose** or **artificial nose**. This thesis focuses on this class of instruments.

1.2.1 A bio-inspired concept

An electronic nose (eNose) takes from the natural olfaction only the main steps which are well-known [PD82; GB94] and which have been detailed earlier in Introduction. First, the transport of the VOC by respiration to the olfactory epithelium is simply replaced by a pump.

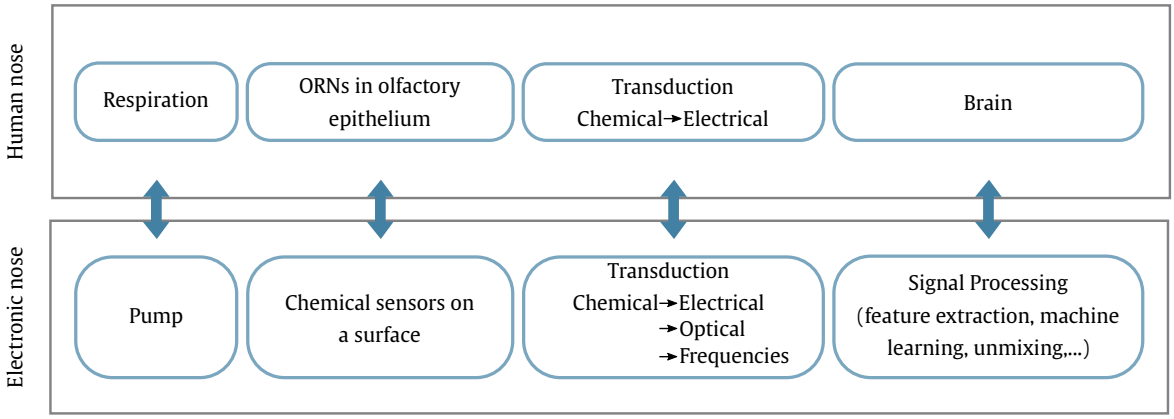


Fig. 1.6.: An electronic nose takes inspiration from the main steps of natural olfaction.

Instead of carrying the VOC to the epithelium, the pump carries it to an array of several different **chemical sensors**. A chemical sensor is mainly composed of a **sensing material** capable of interacting with the VOC, so it plays the role of an ORN [Boe14]. Chemical sensors used by eNoses are mainly **non-specific** and can interact **reversibly** with a broad range of molecules. The chemical information resulting from the interaction between the VOC and the chemical sensor is then translated into exploitable information: this is the **transduction method**. The transduction method varies from one technology to another but the nature of the information can be electrical, optical, frequency etc... Instead of having a brain for processing the data coming from the ORNs, eNose relies on **Signal Processing** (including filtering, feature extraction, machine learning, unmixing, ...) for extracting information aiming at identifying and/or quantifying VOCs. Figure 1.6 summarizes the comparison between artificial olfaction with an electronic nose and natural olfaction.

1.2.2 General response curve

Whatever the nature of the response (electrical, optical, etc...), the response of each chemical sensor is a time series displaying the interaction over time between the chemical sensor and the VOC.

This time series is usually obtained using a method called "**3-phase sampling**". First, the chemical sensor is exposed to a reference gas, usually the ambient air (but it could be a more controlled reference), during a certain amount of time: this is the **baseline** acquisition. Second, the VOC (or the gas mixture) is injected, either using a gas bottle or the headspace[3] of a vial: this is the **VOC injection**. During VOC injection, chemical interaction between the VOC and the sensing material occurs. This interaction is not an abrupt phase transition, like activated/not activated or 0/1 states. The interaction is really time-related: the longer the injection is, the greater the response. However, the response does not go to infinity if we inject the VOC infinitely: the response ends up stabilizing around a **steady-state value**. This steady-state value reflects the chemical equilibrium of the chemical reaction which is measured. In fact, the chemical reaction is always reversible, so whatever the nature of the

[3]The headspace corresponds to the gas phase over a liquid evaporating in a sealed vial.

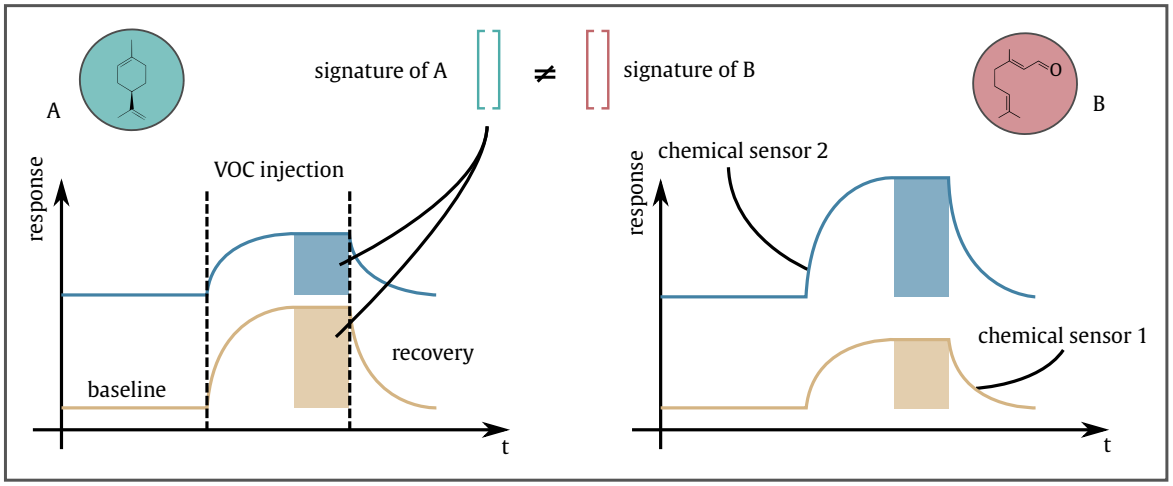


Fig. 1.7.: Data from an electronic nose, for two VOCs (A and B) and two chemical sensors (yellow and blue curves). A classical way to acquire a measurement is to follow a 3-phase procedure: baseline (*e.g.* response to air), VOC injection and recovery (*e.g.* reinject air and stop VOC injection). The result is a time series for each chemical sensor highlighting the interaction between the VOC and the sensing material. We generally summarize the time series by extracting one or two values (for instance, the response at equilibrium) to build a signature of the VOC. Ideally, two different VOCs have different signatures, so we can discriminate them and tell which one is A or B.

reaction, there are two possible events: it can go from one side to the other and *vice-versa*. When the response reaches the steady-state value, it means that these two events became sort of equiprobable, causing the response to no longer change. The time that the chemical sensor needs to reach equilibrium is often referred as the **response time**. This response time can vary greatly from one technology to another (from several seconds to several minutes). Then comes the third phase: the **recovery**. During recovery phase, the VOC injection is stopped and the reference gas is again injected. Since there is no longer VOC supply, the chemical reaction can then only go from one side to the other and the response eventually comes back to its initial value. The time the chemical sensor needs to come back to the initial value is called the **recovery time**, and this time varies also from one technology to another. In Figure 1.7, we report the classical curves obtained after a 3-phase sampling.

The time series obtained result from the reaction between the VOC and the array of chemical sensors. The nature of the reaction (*e.g.* binding, reduction-oxidation, ...) depends on the technology used. Whatever the nature of the reaction, the chemical parameters driving the reaction is both dependent on the VOC and on the chemical sensor. So, different chemical sensors will lead to different chemical parameters and thus different chemical reactions (but of the same nature). Alone, the use of a chemical sensor is limited, however, several chemical sensors carry different information about the VOC due to their different affinity. We say that the eNose creates a chemical "**signature**" of the VOC. A common step to extract this signature from the time series of the chemical sensors is to extract only the steady-state value of each time series: this is what the literature usually calls the signature. The extraction of the signature from the time series is therefore just a feature extraction step, and other features can obviously be used (in particular, dynamics parameters [Yan+15]). The signature

can be seen as a vector living in a space of dimension the number of chemical sensors, and summarizing the chemical interactions. This signature is dependent on the VOC but also on its concentration and many other parameters such as humidity, temperature, age of the sensors,...

Two different compounds, say A and B, lead to different chemical reactions with the sensor array. Therefore, the time series will be also different and likewise for the signatures. In practice, several measurements are carried out with these two VOCs and signatures are extracted from all these experiments in order to build a data base. **Machine Learning** algorithms are then trained using this data base in order to classify future samples as A or B. **Regressions** can also be performed in order to estimate their concentration, based on calibrating samples. Figure 1.7 reports the extraction of the signatures from the time series of an eNose-based system.

1.2.3 Why non-specific sensors?

To understand the use of non-specific sensors for eNoses, it helps to understand the properties that a chemical sensor must, at least partially, observe [BS10]:

- **Sensitivity**: change of response per VOC concentration unit.
- **Limit of detection**: lowest concentration of the VOC that can be detected (related to sensitivity).
- **Selectivity**: characteristic of the sensor to respond to a given VOC in the presence of interferent VOCs.
- **Stability**: reproducibility of the response over time.
- **Response time**: time to reach the steady-state value, starting from the reference value.
- **Recovery time**: time to return to the reference value after an injection.
- **Reversibility**: the sensing material returns after a measurement to its initial condition.
- **Dynamic range**: range of the VOC concentration between the limit of detection and the highest possible concentration.

An ideal chemical sensor would be highly-selective and highly-sensitive with a low limit of detection but large dynamic range [HGO08]. In addition, it would have fast response and recovery times and would show a perfect reversibility [HGO08]. Finally, an ideal chemical sensor would also be stable over a long period of time [HGO08].

Unfortunately, a chemical sensor cannot fill all these requirements, so an ideal chemical sensor does not exist [BS10]. In fact, some of these properties impose contradictory constraints. For instance, it is possible to design highly-selective and highly-sensitive chemical sensors towards a targeted VOC but these properties impose strong interactions between the VOC and the sensing material. In this way, the sensing material will not be sensitive to any other VOCs which could act as interferents in the response. However, these strong interactions will generally prevent reversibility, leading to a short lifetime, a slow recovery time and a lack of stability over long operation time [HGO08]. In fact, reversibility implies weak interactions

but weak interactions cannot lead to high-selectivity. Consequently, chemical sensors will satisfy only few of these properties depending on their type and on the application.

For eNoses, non-specific sensors are used to ensure reversibility but at the cost of a partial selectivity and of a cross-sensitivity to other VOCs than the VOC of interest. The combination of several sensors can then help to cope with these drawbacks.

1.2.4 Existing technologies

The field of eNoses is a 40-year-old field, dating back to the paper of K. Persaud and G. Dodd [PD82]. Ever since then, several technologies have been investigated to be used as part of an eNose system. In this section, we detail the main technologies that can be classified according to the nature of the physical change caused by the chemical interaction: conductivity, mass, optical properties. We also mention the commercial availibility of the technology.

Conductivity-based sensors. With conductivity-based sensors, the interaction between the sensing material and the VOC leads to a change in resistance.

The most used kind of conductivity sensors is undubitably **Metal Oxide Semiconductor** sensors (MOS or MOX). MOS sensors are based on a reduction-oxidation reaction between the VOC and the sensing material, a metal oxide (combination of a metal with oxygen atoms). Reduction-oxidation reactions are based on electrons transfer: in a reduction reaction, the reactent takes one or several electrons whereas in an oxidation reaction, the reactent loses one or several electrons. A VOC is said to be reducing if it can give one or several electrons to a reactent, and oxidizing if it can take one or several electrons. There are two types of semiconductors: n-type (SnO_2, ZnO, ...) reacting mainly with reducing gases (CH_4, CO, ...) and p-type (CuO, NiO, ...) reacting mainly with oxidizing gases (O_2, NO_2, ...). n-type sensors are often preferred for their stability [Pea+06] so we detail only the working principle of n-type sensors. At room conditions, oxygen atoms present in the air are naturally adsorbed[4] on the surface of the metal oxide, forming a layer. When the metal oxide is heated at a high temperature (>160°C), this layer of oxygen atoms ionizes, forming a layer of oxygen ions. When a reducing VOC is carried over the sensing surface, the VOC reacts with the adsorbed oxygen ions present at the surface. This

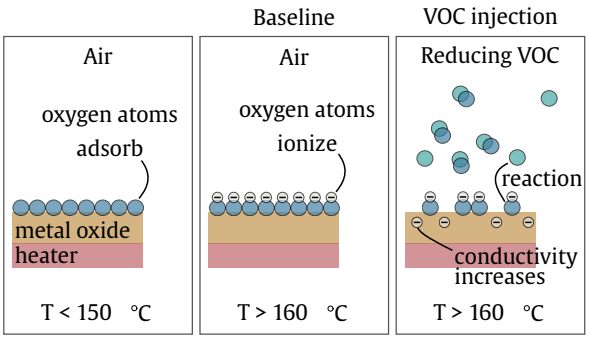


Fig. 1.8.: Working principle of a MOS-based chemical sensor (n-type with a reducing VOC). First, oxygen atoms adsorb on the metal oxide at room conditions. Then, the metal oxide is heated causing the ionization of oxygen atoms. Finally, when a reducing VOC is injected, some oxygen atoms are captured by the VOC causing the release of their electron into the surface, thus increasing the conductivity.

[4] A**d**sorption ≠ A**b**sorption. Adsorption is a surface phenomenon: the molecules (or ions, atoms) adhere to the surface of an adsorbent. Absorption is an assimilation phenomenon: the molecules enter the volume (the bulk) of another substance (the absorbent).

VOC gains an oxygen and releases the electron to the surface: the number of free electrons increases. This increase in the number of free electrons increases the conductivity of the sensing material which results in a decrease of the resistance. If the VOC is rather oxydizing, the reaction is inversed and conductivity of the sensing material decreases (resistance increases). For a p-type sensor, the mechanism is similar but of opposite sign [Ste78; Pea+06]. Figure 1.8 summarizes the working principle of a MOS-based chemical sensor and a good explanatory video can be found on the website of Figaro[5] company (Japan). This type of sensors has been the first chemical sensors used for an eNose [PD82] and several companies have been created since then (*e.g.* Figaro[5], Alpha MOS[6]).

Another kind of conductivity-based sensors is **Conducting Polymers** (CP). As the name suggests, CP is a compound which can conduct electricity. A polymer is a long chain formed by the repetition of a molecule, called a monomer. When the polymer adsorbs some VOCs, the adsorption changes its conducting property and this change can be measured as a resistance change. Several polymers have been studied, different polymers leading sometimes to different processes [Ars+04], in particular polypyrrole-based composites and carbon black based composites. As for MOS-based sensors, the use of CPs in an eNose has been intensively studied and CPs are currently used in the eNose of Sensigent[7] company.

Piezoelectric sensors. Piezoelectric materials are materials which generate electricity in response to a mechanical stress and can be mechanically stimulated in response to an electrical field. For this kind of sensor, adsorption of the VOC on the sensing material leads to a change of mass which is measured by a shift in resonant frequency.

An example of piezoelectric sensors is **Quartz Crystal Microbalance** (QCM) sensors. Quartz is a piezoelectric material composed of silicon dioxide (SiO_2) which can oscillate at its resonant frequency (10-30 MHz) when it is exposed to an electrical field [SBE98]. In the case of QCM, this electrical field is generated by applying an electric potential by means of two electrodes placed on both sides of a disc-shaped quartz crystal. The quartz is coated with a sensing material on which VOC can adsorb. The adsorption of the VOC on the sensing material results in a change of mass which in turn leads to a change of the resonant frequency [Ars+04]. Figure 1.9 reports the working principle of a QCM-based sensor.

Surface Acoustic Wave (SAW) sensors work according to the same principle. They use a plate of piezoelectric material (ZnO or lithium niobate) and interdigitated electrodes (*i.e.* comb-like electrodes which are interlocked) are put at two ends. In between the electrodes, the piezoelectric material is covered with a sensing material. From one side, a set of electrodes excites the piezoelectric material with an electrical field, resulting in a surface wave propagating along the piezoelectric material including the sensing material and reaching the other side. On the other side, the other set of electrodes simply converts

[5] http://www.figarosensor.com/movie/ (March 2, 2020).
[6] https://www.alpha-mos.com/ (March 2, 2020).
[7] https://www.sensigent.com (March 2, 2020).

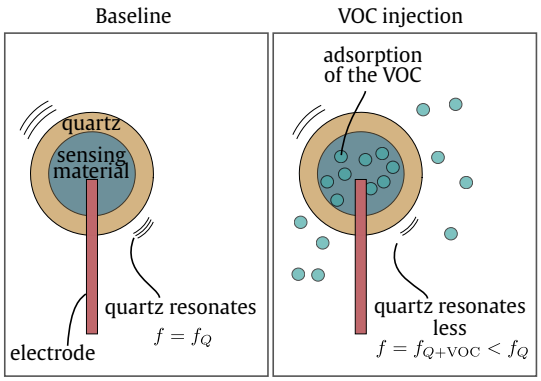


Fig. 1.9.: Working principle of a QCM-based chemical sensor. A quartz is coated with a sensing material. When the VOC adsorbs on the material, it changes the resonant frequency.

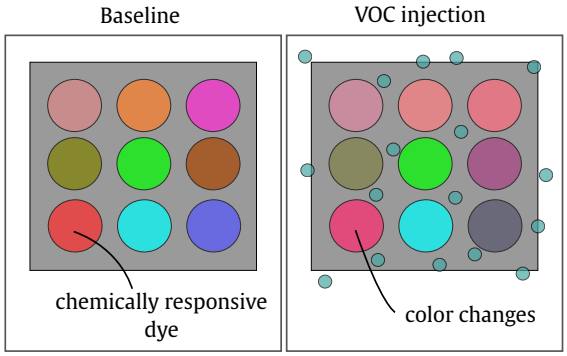


Fig. 1.10.: Working principle of a colorimetric chemical sensor. The color of a chemically responsive dye changes in presence of a VOC.

the mechanical stress (due to the acoustic wave generated by the first electrodes) into an electrical measurement. Again, the adsorption of a VOC on the sensing material causes a frequency shift of the surface wave [Pea+06].

Optical sensors. The interaction between the VOC and the sensing material can sometimes lead to a change in optical properties. One of the main advantages of this transduction method compared to the previous ones is that the increase in the number of sensing materials does not make the architecture of the system more complex.

Colorimetric sensors use a change in color to measure chemical interaction [RS00]. For that, it consists in using chemically responsive dyes (*e.g.* metalloporphyrins). When the VOC binds to one of these dyes, it changes the absorbance spectrum of the dye. A simple RGB camera is then used and a color change reports the interaction. One of the most advanced works with colorimetric sensors is from Suslick's group [RS00], which resulted in a commercial eNose sold by iSense[8] company (United States). Figure 1.10 reports the working principle of a colorimetric sensor.

Another category of sensors based on optical methods are **Surface Plasmon Resonance** (SPR) based sensors. This class of instruments will be introduced in the next section. In fact, the electronic nose used in this thesis is based on SPR imaging.

To conclude, there exists a lot of different technologies and more are detailed in several reviews [Alb+00; Ars+04; RBW08; Lak+14]. Sometimes, a comparison table is reported by the authors of the reviews. In fact, there exists no study benchmarking all the available technologies of electronic noses for a given application or a given set of VOCs. So, comparison is limited since it can only be done by comparing different studies of different authors using different VOCs with different methodologies and different aims. To be clearer, the comparison could be done if we only compare the size, the cost or the stability of the final device. However,

[8] http://isensesystems.com/ (March 2, 2020).

a comparison of sensitivity, response time, limit of detection or selectivity would depend on the targeted VOCs.

1.2.5 Challenges

Whatever the technology used, several challenges for Signal Processing result from the use of non-specific chemical sensors by an electronic nose.

An eNose has basically two goals: identification and quantification. However, in comparison to highly-specific chemical sensors, the two tasks are not straightforward at all. In fact, highly-specific sensors respond theoretically only to one VOC whatever the other VOCs present at the time of the measurement. If it reacts, it means that the VOC is present [Boe14] and it is over for the identification part. For quantification, this kind of sensors are often made in order to respond linearly over a range of concentration which is as large as possible. So, quantification is also direct if we have some samples for which we know exactly the concentration.

For non-specific sensors, it is not that simple, even if the measured VOC is pure. If the VOC is pure, it is unlikely that only one chemical sensor responds while the others remain at zero. Due to the non-specificity of each chemical sensor, a large part of chemical sensors, if not all, can respond positively to the VOC. So, the identification is no longer straightforward since all the chemical sensors would also respond to another VOC.

Let us review all the issues step-by-step. Raw data is a collection of P time series (P the number of chemical sensors), and the first issue is about extracting one or two values from each time series: it refers to **feature extraction** (*e.g.* which information will summarize the time series?). Recall that each chemical sensor reacts with the VOC depending on its affinity with it, and each chemical sensor is expected to have a different affinity, so the variations in affinity across chemical sensors are expected to be specific to the VOC, making possible its identification. However, two signatures can look alike, due to tiny differences in affinity. Even with stronger differences, comparing by sight two vectors of high dimension (several dozens), is not an easy human task. So, it calls for **Machine Learning** algorithms, and more precisely **classification** algorithms (clustering could also be considered, but since we are more interested in identifying the VOC rather than in finding groups in data, clustering is of little interest). Any classification algorithm requires a training set, relating signatures to labels (VOC identity), so, any task starts by carrying a lot of **experiments** measuring a lot of times several VOCs. Ideally, these experiments are carried out under real-life conditions. Under such conditions, the VOC concentration cannot be reliably controlled. The chemical reactions are dependent on the VOC concentration, so the time series are also dependent on the VOC concentration and the same holds for the resulting signature. It could cause problems for identification. If a classifier learns to differentiate A from B in a given range of concentration, it is unlikely that it will generalize out of the range since the signatures will not be the same. So **normalization** of the signature is often first applied to remove variations in concentration in the signature. After normalization, ideally, only the variations in affinity remain. Once the signatures have been normalized, **dimensionality reduction** techniques

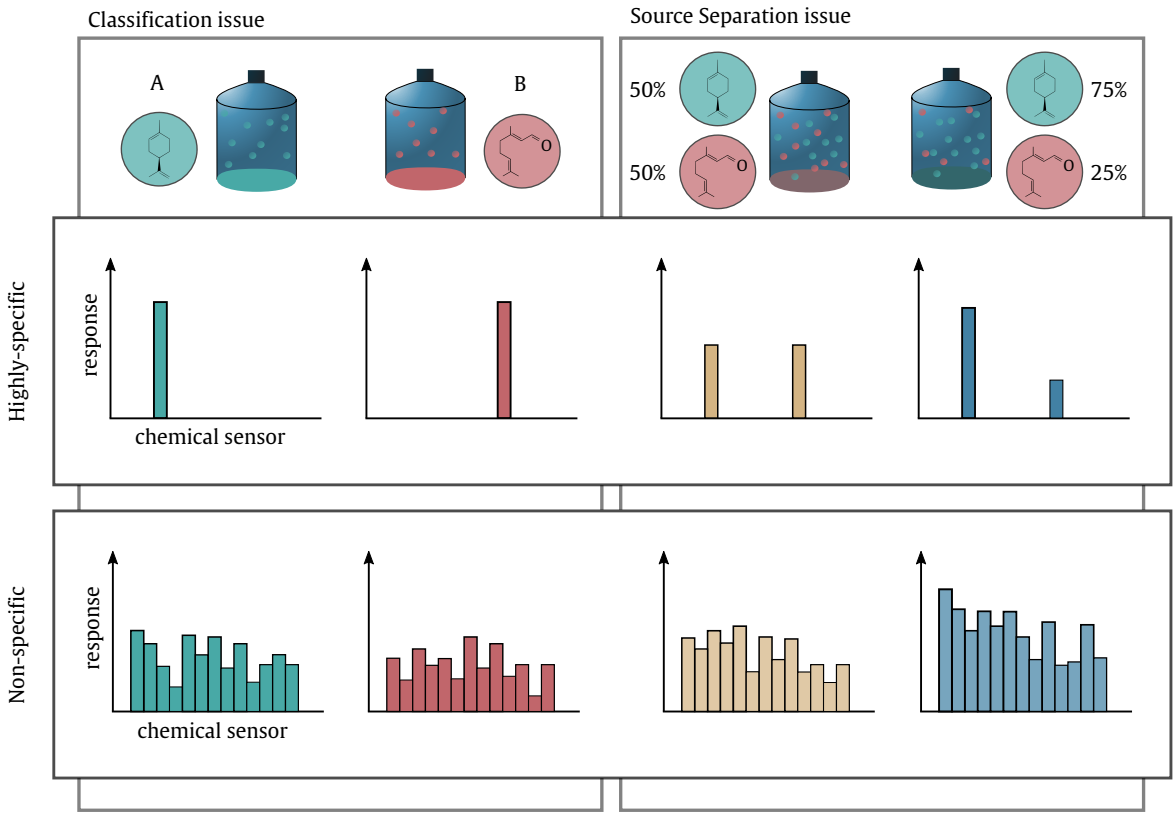


Fig. 1.11.: Classification and Source Separation issues with an eNose (non-specific sensors) compared to the use of highly-specific sensors. Signatures are obtained after a feature extraction step from the time series and are represented as histograms (each bar corresponds to a response of one chemical sensor to the VOC, characterizing the interaction between the VOC and the sensing material). First row corresponds to signatures obtained with highly-specific sensors while second row corresponds to signatures obtained with non-specific sensors (used by an eNose). Figures on the left panel illustrates the classification problem (for two pure VOCs A and B). For highly-specific sensors, the task is straightforward while the use of non-specific sensors calls for Machine Learning algorithms. Figures on the right panel illustrates the Source Separation issue for two different mixtures of A and B (50% A & 50% B; 75% A & 25% B). Again, the task is straightforward for highly-specific sensors: the amplitude of each individual sensor gives us the concentration of A or B in the mixture. With non-specific sensors, estimating concentrations is not trivial and calls for Source Separation algorithms.

are used to, as the name suggests, reduce the number of dimensions (at least equal to the number of chemical sensors). Then, a chosen **classifier** is trained from the training set and tries to predict the labels of future samples. If the pure VOC has been identified, **modeling** and **regression** (linear or not) are used to estimate its concentration.

However, if the VOC is not pure but a mixture of several different VOCs, all the previous steps become even harder. Once again, for highly-specific sensors, that would not be so hard: if two VOCs, say A and B, are mixing and are injected in the instrument, only the two chemical sensors related to A and B will respond positively, and again, it is over. For non-specific sensors, each chemical sensor can presumably interact with the two VOCs at the same time, due to their cross-sensitivity. So, the response of the chemical sensor will not depend only on A or on B but on both. In fact, the instrument will generate a signature for the mixture which is not directly the signature of A or the signature of B but another signature which is related

to A and B in a non-trivial way, depending on the proportion of each VOC in the mixture. In the literature, some authors deal with this case as another classification task where a classifier learns different proportion of A and B (10% of A & 90% of B, 50% of A & 50% of B, etc... See for instance [BOD12; Ber+12]). In the same vein, other authors propose a learning approach: we call this approach mixtures learning. The idea is to generate plenty of mixtures of A and B at different known concentrations. Then, a learning algorithm (generally a black-box algorithm based on neural networks) tries to learn the model relating the signatures to the concentrations (see for instance [Fon+15]). The fundamental issue of mixtures learning is that it requires the generation of the mixtures to learn the model: regarding experimental practicability, it becomes rapidly unmanageable if we assume more than two VOCs. Mixtures learning will be more discussed in Chapter 6. In fact, gas mixtures naturally call for **Source Separation** methods (sometimes called **Unmixing**). Figure 1.11 illustrates the issues related to Source Separation and classification.

The previous paragraphs neglect some drawbacks of chemical sensors, which can be detrimental for all the challenges described above. First, chemical sensors are sensitive to environmental conditions. This sensitivity can vary from one technology to another but **temperature** and **humidity** will often play key roles in the response. Almost any chemical reaction is affected by the temperature, and humidity acts as an interferent in the measurement. In fact, most technologies will respond positively to water vapour. Second, even if these environmental conditions are controlled or their contribution to the response removed, any instrument based on chemical reactions will **drift** over time. Drift is a dynamical and time-related process which affects the reproducibility and repeatability of the instrument. Concretely, the signature of a VOC, even in the same environmental conditions, will change over time, causing a serious and obvious problem for classification: the classifier will not generalize to future samples if too much time has elapsed since the training set was acquired [Hol+97]. Drift can be caused, *e.g.*, by sensor ageing or by irreversible binding reactions on the sensing material. It is the topic of Chapter 4 in this thesis. Note that drift also occurs in well-established analytical instruments, such as GC-MS [RP+18].

To sum up, an eNose is challenging regarding several topics related to Signal Processing: classification (feature extraction, dimensionality reduction), modeling, quantification, source separation and drift.

1.2.6 A short critical review of the literature

A specific literature will be introduced chapter-by-chapter, focusing on the themes of each chapter. eNose methodology has been criticized over the years by several authors [Mie96; MML00; Boe14; Mar14]. These criticisms concern issues related to statistical reliability, experimental confounds and even semantics [Boe14]. In this section, we point some critical aspects out, especially data-related aspects.

A small number of sensors. The number of chemical sensors used by most studies in the field is small, in the range of a dozen[9] (counting the replicas). If we compare to the human nose, having several millions of ORNs out of ∼400 different types, most of the electronic noses are far behind.

To illustrate this small number, Table 1.1 reports the number of sensors used by the authors of the top 10 of the most cited articles in the past 5 years. Table 1.1 highlights a typical number of sensors between 6 and 12 and also demonstrates that the literature is dominated by one technology: MOS-based sensors.

Ref.	Technology	Number of sensors
[Fon+16]	MOS	8 (4)
[Sai+18]	MOS	6 (6)
[Khu+17]	Colorimetric	12 (12)
[LGW17]	QCM	8 (8)
[NC+17]	MOS	6 (6)
[HB16a]	MOS	8 (8)
[Jia+17]	MOS	10 (10)
[MGJ17]	MOS	10 (8)
[HB16b]	MOS	7 (7)

Tab. 1.1.: Number of sensors used by the top 10 (on February 25, 2020) of the most cited articles in the past 5 years according to the Scopus data base, by limiting to *Sensors* and *Sensors and Actuators B: Chemical* journals. In parentheses is the actual number of sensors (removing the replicas).

The question "*Are more sensors better?*" (quoted from A. Hierlemann and R. Gutierrez-Osuna [HGO08]) is a controversial question in the eNose field. First, all authors agree with the fact that the use of redundant chemical sensors or replicas (*i.e.* using the same sensing material) is beneficial. It is well known in the Signal Processing community that, averaging the outputs of J redundant sensors allows the variance of the noise to be reduced by a factor J compared to an individual sensor ($\sigma^2_{\text{aver}} = \frac{\sigma^2_{\text{ind}}}{J}$). However, there is a debate regarding the interest of increasing the diversity of the sensor array with different sensing materials.

On the one side, multiple authors claim that such large sensor array could have many disadvantages. First, for a lot of technologies, increasing the number of sensors can considerably increase the complexity of the system and its cost (*e.g.* the size, the power consumption in case of MOS-based systems, ...) [HGO08]. In this case, it may be more reasonable to think about an optimal number of sensing materials rather than always increase this number. But this "optimal" number of sensors is a quite difficult question in practice. Second, increasing the number of sensors increases the dimensionality of the problem, leading to the so-called curse of dimensionality[10]. For instance, increasing the dimensionality while keeping the

[9] There are some exceptions such as the eNose from iSense[8] company which is, to our knowledge, the actual largest commercial eNose since it can board up to 120 colorimetric chemical sensors.

[10] The curse of dimensionality was introduced by R. E. Bellman [Bel57; Bel61]. A famous example of the curse of dimensionality is a problem related to sampling. Suppose that one wants to sample a space with a spacing of 0.01. In a 1-dimensional space and limiting to the unit interval ($[0, 1]$), it corresponds to 100 samples which is quite reasonable. However, in a 10-dimensional space, with the same spacing and limiting to the unit interval for each dimension, a simple calculation leads to $(10^2)^{10} = 10^{20}$ samples, which is much less reasonable. This example illustrates the exponential increase in volume due to an increase in the number of dimensions. Another famous example, introduced by K. Beyer et al. [Bey+99], concerns a problem related to distances between samples. As dimensionality increases, the distance between a sample and its closest

number of samples constant can decrease classification performance [Hug68]. Third, adding sensing materials does not mean that these sensing materials are relevant for the sensing tasks and can even be sensitive to irrelevant interferents due to the non-specificity of the sensors [JRP15]. They can even add noisy measurements leading to a decrease in performance [Pea+06]. Finally, small array of sensors can already discriminate a significant amount of different compounds [SP02].

On the other side, authors claim that the more, the better. For several authors, the idea of a small array of sensors as being sufficient can hold only in simplistic scenarios where the environment is controlled and pure VOCs at known and fixed concentrations are discriminated [Alb+00]. In practical applications, VOCs of interest are rarely pure, often present in mixtures and polluted by interferents [Alb+00]. The increase in the number of sensors in an eNose is then no longer seen as a drawback but as a chance to disambiguate similar signatures [Alb+00] while potentially reaching the performance of its biological counterpart [Bec+10]. Fears related to the curse of the dimensionality are of course justified. However, by coupling this increase of diversity with appropriate dimensionality reduction methods, the drawbacks of the curse of dimensionality can be mitigated: noisy sensors can be identified, sensors irrelevant to the sensing task can be discarded and informative sensors can be selected. In fact, many fields have to naturally deal with high-dimensional data and all is not lost in high-dimension (*e.g.* principal directions can be consistent when the number of sensors is much greater than the number of samples by imposing sparsity to PCA [JL09]). But more sensors imply often more samples, which will lead us to the next point.

To conclude on the question "*Are more sensors better?*" [HGO08], the two sides may have simply a different vision of an electronic nose. On one side, an eNose can be tuned for a specific application, selecting a set of non-specific sensors having a good sensitivity towards the targeted VOCs. On the other, authors may have a more universal vision of an eNose, as being a broad-based measurement device usable for a large number of various applications. In this thesis, the eNose used will never be specifically tuned towards a given application and is therefore in line with the second definition.

A small number of samples. Broad-based chemical sensors used by an eNose are plagued by drift over time, are noisy and are sensitive to external influences (temperature, humidity, ...). In addition, non-specificity, which is often seen as a strength, is also a weakness since the VOC (or the VOCs) responsible for the measure is not well defined [Boe14]. All these factors produce a great measurement uncertainty which calls for prudence regarding overoptimistic conclusions which are often not supported by substantial data sets [GDR01; Boe14] and lack of validation procedures [Mar14] (for instance, repeating the experiment over separated measurement sessions). The small-dataset issue seems particularly true for food-related [RBW08] and health-related studies [Mar14], where it is not rare to find classes with 4-5 samples. Despite this small number, some authors state conclusions such that "*This paper*

neighbor becomes equivalent to its distance to its farthest neighbor. This counter-intuitive effect can occur for as few as 10-15 dimensions.

clearly shows that it is possible to differentiate and classify red wines coming from the same cellar as well as from the same variety of grapes and same geographic origin".

The problem with small datasets concerns the confidence in the results. This topic can be easily found in any good book of Pattern Recognition (see [Mit97] or [DHS12]). Imagine that we want to show that an eNose is capable of discriminating two VOCs, say A and B, for a given classifier. This ability of the classifier to differentiate A from B is represented by the true classification rate, say $\mathcal{P}$. This rate tells us the probability to correctly classify as A or B a new sample taken randomly from the data distribution. If $\mathcal{P}$ is far from the chance level (here 50% since we have 2 classes) then we can conclude that the eNose is able of differentiating A from B. However, $\mathcal{P}$ is an unknown quantity so we need to estimate it in practice: this is the estimated classification rate $\hat{\mathcal{P}}$. This estimation requires data, so we carry out several experiments to generate several samples, for A and B. Say that we have acquired a total number of N samples, with an equal proportion of A and B (so $\frac{N}{2}$ samples/class). The acquisition of these N samples is assumed to be independently drawn from the data distribution. Then, a standard procedure to estimate $\mathcal{P}$ is to use k-**fold cross-validation** (CV). In a k-fold CV, the data set is randomly split into k groups: $k-1$ groups are used to tune the parameters of the classifier (it is the training set) and, using these parameters, the k^{th} group is classified (it is the test set). This procedure is repeated for each one of the k groups. Finally, the N predicted classes are used to compute $\hat{\mathcal{P}}$, by comparing predictions and ground truth (we know the class of each sample). However, the estimated classification rate results from a random sampling of the data distribution, so this estimation is affected by uncertainty. A resulting question is then, how good that estimation is?

The answer relates to **confidence intervals** which relate in turn to the number of samples N. Confidence intervals are a range of rates which are estimated from the data (so they are random too) and where we would expect to find the true rate. It is always connected to a confidence level, generally a confidence level of 95%. A 95% confidence interval is an interval that is expected with probability 95% to contain the true rate $\mathcal{P}$ [Mit97]. More precisely, if we could repeat over and over again the same experiments with the same number N of samples, then 95% of the generated confidence intervals would actually contain the true rate. How can we compute confidence intervals? Let us simplify the problem by considering that all the test samples are independent. This is a strong assumption which is often violated in practice since one usually uses cross-validation procedure, making the test samples dependent (for two different test sets using CV, the corresponding training sets greatly overlap). However, the assumption of independent test samples makes the theory simpler and is useful to understand the problem. In theory, we can see the estimate $\hat{\mathcal{P}}$ as the result of a series of binary decisions: the sample is well classified or not. In Statistics, such a binary decision is called a Bernouilli trial, following a Bernouilli distribution of parameter $\mathcal{P}$ (the true probability to correctly classify a new sample). The series of N independent binary decisions follows a Binomial

distribution[11] of parameters N and $\mathcal{P}$ (hence the need of independent test samples). Without going into calculation details, for $N \geqslant 30$, we can use the Central Limit Theorem and the 95% confidence intervals are calculated according to the normal approximation interval:

$$\hat{\mathcal{P}} \pm 1.96\sqrt{\frac{\hat{\mathcal{P}}(1-\hat{\mathcal{P}})}{N}}$$

Figure 1.12 is a figure retaken from the book of Duda & Hart [DHS12] which represents the relation between the number of samples N and the 95% confidence interval.

For instance, to be reasonably sure that the true classification rate $\mathcal{P}$ is greater than 98%, it would require that the classifier makes no error out of 250 test samples [DHS12]. From the same figure, we can also notice that the notion of chance level depends on N [MP+08]. For two classes, a chance level of 50% is in fact an asymptotic rate, which can hold only for large data sets. For small data sets such as $N = 10$, the 95% confidence interval of chance rate is $50\% \pm 26.2\%$ (since N is too small in this case for normal approximation, we used the adjusted interval proposed by Agresti and Coull [AC98]). So the upper bound is 76.2% and any estimated classification rate below or close to this value must be considered with caution. In Figure 1.13, we represent graphically the loss in confidence induced by small data sets.

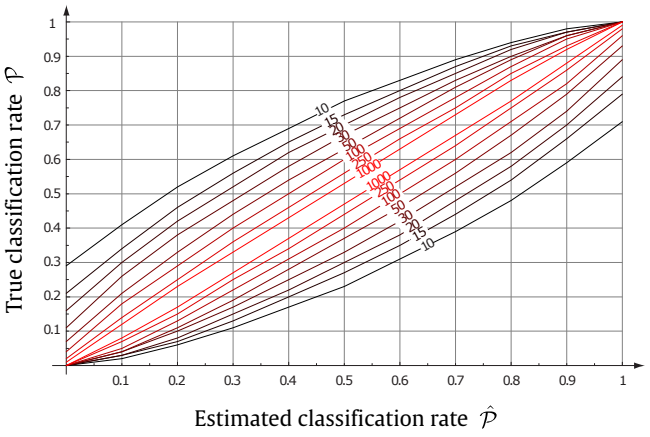


Fig. 1.12.: 95% confidence interval for a given estimated classification rate $\hat{\mathcal{P}}$. For each $\hat{\mathcal{P}}$, there is 95% of chance that the true $\mathcal{P}$ lies between the black and red curves. Numbers near black curves indicate the number of test samples. Figure retaken from [DHS12] (axis labels have been changed to follow our notation).

The 95% confidence interval that we have just introduced is quite interesting to be aware of the confidence problem with small data sets. However, in the eNose field, even this 95% confidence interval can be over-confident. In fact, this interval implies that the samples have been drawn independently and from the same data distribution. Independence is not easy to obtain with real data sets [Mar14] and even more crucial, samples are acquired over time. As we have seen in Section 1.2.5, eNose is prone to drift, which means that data distribution changes over time, so it is likely that time-related samples have not be drawn from the same distribution. This last point calls even more for prudence regarding the results. In fact, the low statistical reliability of eNose studies can incidentally explains why there are so many successful applications in lab environment but so few beyond the lab [Mar14].

[11] The Binomial distribution gives the probability of observing r heads in a series of N independent coin tosses, if the probability of heads in a single toss is $\mathcal{P}$ [Mit97]: $P(r) = \binom{N}{k}\mathcal{P}^r(1-\mathcal{P})^{N-r}$

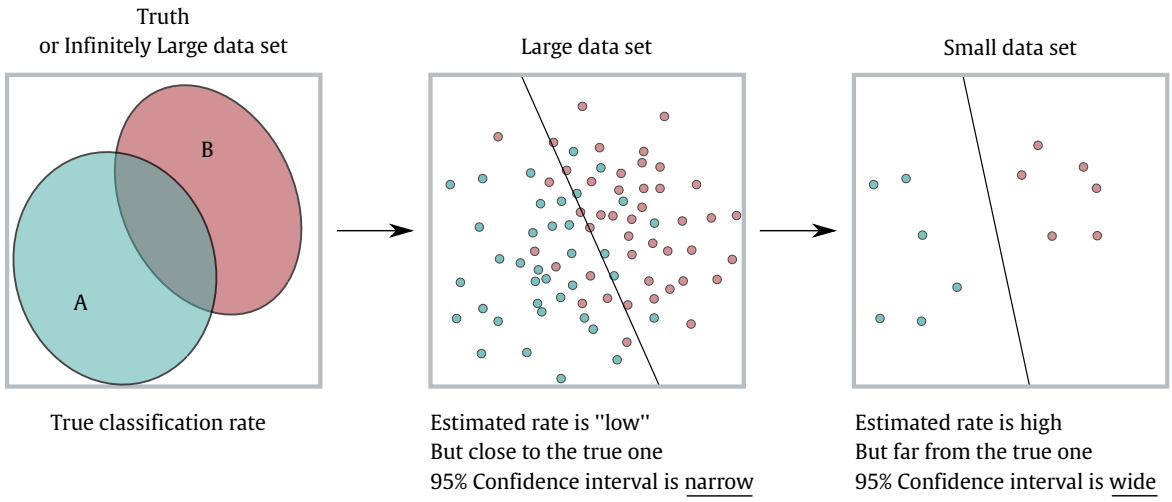


Fig. 1.13.: A small data set leads to low confidence in the results. Black line corresponds to the estimated decision boundary.

A solution is to increase the number of samples. We agree that it is hard to determine the minimum number of samples required to be statistically reliable while being realistic in practice (it is especially true for health-related data where a sample is generally a human and each new sample has a huge experimental cost). However, in Pattern Recognition, a good rule to follow is to have at least 5-10 times more training samples per class than the number of features used by the classifier [JC82; JDM00]. For instance, after a reduction to 5 dimensions, this rule would correspond to have at least 25 samples per class for the training phase. It has also to be noted that this number can increase with the complexity of the classifier [JDM00].

In some studies, authors qualified as a sample each time point of the time series shown in Figure 1.7 (see for instance [Bru+07]). With this definition of a sample, the size of the data set can be quite high even if the number of repeated experiments is quite small. Having good temporal resolution and using dynamical features from the time series may help achieve reliable discrimination. However, this increase can also lead to over-optimistic results: using N time points of a single measurement is clearly not equivalent to generating N independent measurements. Therefore, results obtained by considering the entire time series must be cautious and authors must emphasize the real number of independent samples.

To conclude on this point, small data sets are common in the eNose field [Boe14] whereas the measurement uncertainty is high. This often leads to over-confident results. Following Pattern Recognition recommendations [JC82; JDM00], it is a good practice to have at least 5-10 times more experiments per VOC than the number of features actually used by the classifier.

An abusive terminology. Earlier in Introduction, we have compared the stages of an eNose-based system to the stages of its biological counterpart. The terminology “electronic nose”, “artificial nose”, “artificial olfaction” can then be tolerated to draw attention to these parallels.

However, the comparison stops there and an eNose is clearly not the direct electronic version of a human nose. To quote J. R. Stetter and W. R. Penrose [SP02], "*eNose devices respond to the chemicals to which the sensors respond. The concept of odor is a human one and so human receptors/sensors respond sometimes to totally different compounds than the eNose sensors.*". Therefore, talking about "odor", "smell", "scent", "odorant", "fingerprint" or "smellprint" (for the signature of a VOC) is quite misleading[12]. According to the Collins definition, "*the smell of something is a quality it has which you become aware of when you breathe in through your nose*". So, a smell is highly related to perception but an eNose does not perceive anything. An eNose only reports chemical interaction between chemical sensors and a VOC (or a gas mixture) [SP02]. For illustrating this fact, there are plenty of examples. For instance, an eNose can detect carbon monoxide while the human nose cannot. In addition, smell and taste are also highly influenced by other senses and can be misled by them [Spe15]. A good example is the study of Morrot et al. [MBD01]: in a discrimination task of white and red wines, they succeeded in misleading a human panel just by coloring a white Bordeaux wine with a red odorless food dye. In this case, it is clear that if the dye does not have any response to the eNose, the eNose will not mislead.

A misleading terminology leads sometimes to misleading methodology [Boe14]. To any newcomer in the field, we recommend to read the article of P. Boeker [Boe14] who warned researchers in the field and criticized many aspects of the methodology and terminology used. In fact, almost two decades before him, P. Mielle did same [Mie96; MML00]. However, it is still quite common to find research articles[13] defending eNose-based system as an odor measurement system (*e.g. Bioelectronic nose and its application to smell visualization* [KP16], *Application of electronic nose as a non-invasive technique for odor fingerprinting and detection of bacterial foodborne pathogens: a review* [Bon+19]).

1.3 An SPRi-based optoelectronic nose

The electronic nose used in this thesis is a commercial optoelectronic nose developed by the French company **Aryballe** (Grenoble). Aryballe develops a new class of instruments based on the use of **Surface Plasmon Resonance imaging** as transduction method and **peptides** as sensing materials.

This thesis was in collaboration with Aryballe which provided eNose prototypes. Thierry Livache and Cyril Herrier from Aryballe took part in the thesis and helped us both with the chemical aspects and the characterization of the instrument. Throughout this thesis, the results shown with Aryballe's eNose have been obtained with different eNoses and versions of the eNose, since the device was still in maturation phase. For instance, the number of chemical sensors used and their nature can vary from one Chapter to another. However, the total number is always around 60.

[12] To be honest, my first poster defined an electronic nose as an odor sensor.

[13] Here, we only point researchers' terminology out. In fact, this terminology is also often used by marketers or with a popularization aim. In this case, we agree that "Volatile Organic Compound" can be less appealing than "smell" or "odor".

1.3.1 Short history

The first SPRi used in liquid phase for biological analyses was developed in early 2000 at CREAB[14] and Institut d'Optique d'Orsay, via the startup company Genoptics (now included in Horiba Scientific) [Gue+00]. The gas phase developpement has been carried out since 2013 at CEA, at INAC institute, in SyMMES lab by the CREAB team. In 2014, the process developed by the team was patented and published the next year [Liv+15]. In parallel, a technology transfer and a partnership was concluded between the CEA and Aryballe (at this time, a start-up called Aryballe Technologies). Aryballe successfully developed a miniaturization of the device formulated by the CREAB and started to sell the instrument since 2018.

From 2016 to 2019, a consortium has been established between Aryballe and several partners (industrial and academic) inside the WISE project, supported by the FUI. The CNRS is involved in this project, in particular represented by GIPSA-lab. This thesis is part of this project and is also supported by the BPI.

1.3.2 Working principle

Chemical sensors. Sensing materials are mainly peptides (short proteines composed of less than 10 amino acids), provided by Smartbioscience (France). Some small organic molecules are also used in addition to peptides as sensing materials. The sensing materials are grafted on a gold-covered prism using thiol-gold bonding. To form an array of chemical sensors over the prism surface, Aryballe uses a FlexArray S12 printer (Scienion, Germany) according to a process previously described ([Hou+12], [Bre+18]). Due to the collaboration with Aryballe and for confidentiality reasons, the nature of the sensing materials and the functionalization principle (the way the sensing materials are fixed over the surface) are not detailed in this thesis.

During an acquisition, the VOC is brought above the gold surface by a flow of air using a pump. The VOC can then interact with the sensing material through a reversible binding reaction. This reaction is both dependent on the VOC and on the sensing material. Thus, different sensing materials will lead to different chemical reactions, creating a chemical signature of the VOC. Since different VOCs lead to different chemical reactions, and thus different signatures, we are able to recognize VOCs. In all the experiments reported in this thesis, the nature and the number of chemical sensors used can vary. However, the cross-sensitive chemical sensor array is generally composed of ~20 different sensing materials which are replicated 3 or 4 times on the surface, leading an array of more than 60 elements. In the previous section, we highlighted that the number of sensors generally used by the literature is close to a dozen. So, this number is quite uncommon. In fact, to our knowledge, the actual largest commercial eNose is from iSense[8] company with its colorimetric eNose derived from the work of Suslick's group [RS00], which can board up to the impressive number of 120 chemical sensors. Then, the number of sensors drops radically to a dozen, by way of the eNose of Sensigent[7] company

[14] See "List of acronyms" at the beginning of the thesis.

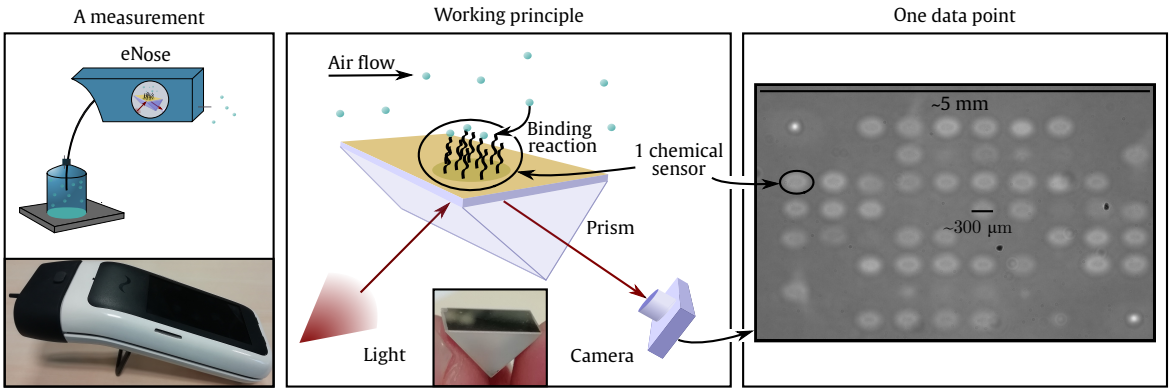


Fig. 1.14.: Optoelectronic nose used in this thesis and provided by Aryballe. Sensing materials are mainly peptides and transduction method is based on Surface Plasmon Resonance imaging (SPRi). Left panel shows an image of the instrument. Middle panel illustrates SPRi: light is sent, reflected by the surface and caught by a simple optical camera. When a VOC adsorbs on the surface, the amount of reflected light increases. A real image of the prism is represented. Right panel corresponds to a real image of the prism surface (one data point).

which sells Cyranose 320 boarding 32 conducting polymers and which has been used in many research studies.

Since the instrument boards peptides, this kind of electronic noses is sometimes called bioelectronic nose [Was+19]. We keep here the denomination electronic nose (and more precisely optoelectronic nose).

Transduction mechanism. The binding reactions at the surface are measured using an optical method, called Surface Plasmon Resonance imaging (SPRi). The physical phenomena behind SPRi will be discussed in the Chapter about modeling (Chapter 3, Section 3.1.2). Briefly, light is sent, reflected by the surface and caught by a simple optical camera. When a binding reaction occurs with the VOC, this changes the refractive index (more light is reflected). The changes in reflectivity are caught by the camera, which thus records in real-time the binding reactions. A representation of the working principle and a real image of the prism surface is presented in Figure 1.14. Light areas correspond to the chemical sensors on the surface.

Since the instrument is based on an optical method, this kind of electronic noses is called optoelectronic nose.

Data extraction. Raw data is a collection of images which reports in real-time the chemical interactions (generally, at a frame rate of 5 Hz). Instead of directly manipulating the images as our raw data (which are quite redundant), we just extract the interesting regions from the images. These regions are the light areas which correspond to the chemical sensors fixed over the gold surface. To identify these areas, a binary mask is manually built before the experiments. Each light area is then summarized as a single value by averaging the pixels, so an image having 60 chemical sensors (60 light areas) produce a vector of dimension 60. This vector is composed of pixel values. To convert these pixel values into reflectivity, we divide

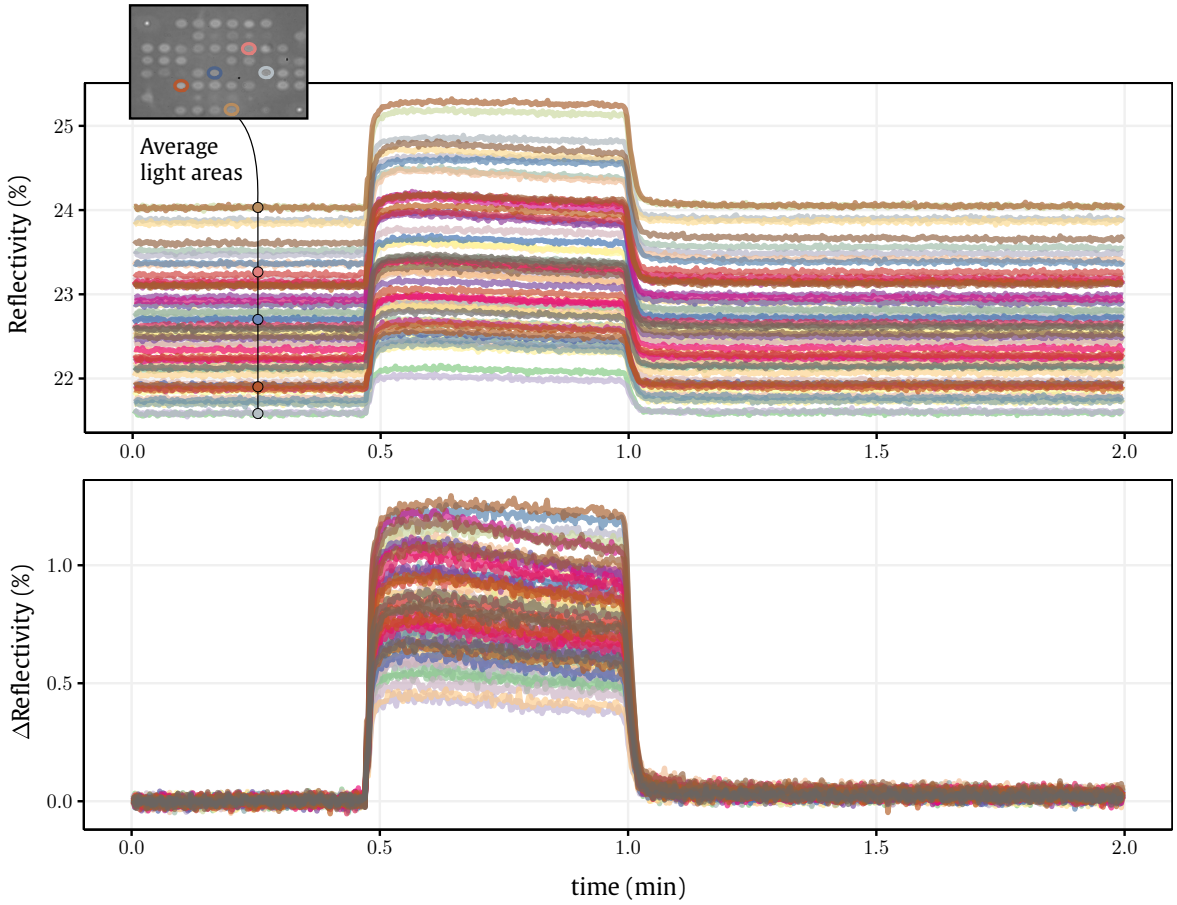


Fig. 1.15.: Data from a 3-phase sampling measurement, before (top) and after (bottom) baseline subtraction. Each line correspond to the time series of one chemical sensor. From one time series, a single time point corresponds to the average pixel of a light area of one image.

by a reference image from which we have extracted the same areas. The acquisition of this image will be later described in Chapter 3, when we detail SPRi.

In Figure 1.15, we report a real measurement according to a 3-phase sampling for a pure VOC ((S)-Limonene). Top figure represents the signals in reflectivity while bottom figure represents the signals after baseline subtraction (so ΔReflectivity). This subtraction removes the reference contribution and it is quite classical in the field to remove the baseline drift (baseline is unstable from one measurement to another). The dynamics is here quite fast and the steady-state values are rapidly reached for the $\sim$60 chemical sensors. In practice, we only keep from these series one or two values per chemical sensor, such as the steady-state value, to generate the signature of the VOC.

1.3.3 Past and contemporary results

Since the patent, several works and theses have been carried out or are currently in progress, especially with a lab version of the Aryballe's eNose developed by the CREAB team. Sophie Brenet worked on the development and the optimization of this lab instrument, showing the first results of the instrument in controlled conditions (for confidentiality reasons, her thesis is not publicly available [Bre18]). Two theses are currently in progress. Charlotte Hurot is

developing new sensing materials for the eNose and Jonathan S. Weerakkody is working on the development and the characterization of the instrument for biomedical applications.

In a recent paper, S. Brenet and coworkers have studied the selectivity and sensitivity of the lab version of the instrument [Bre+18]. First, they showed a good discrimination of both different and similar molecules up to a single carbon atom. Second, they studied the sensitivity of the device regarding two VOCs, namely Ethanol (170-1400 ppm) and 1-Octanol (1.43-5.84 ppm). The concentration range of Ethanol was significantly greater than for 1-Octanol due to a lower affinity of the sensor array to Ethanol. For 1-Octanol, the theoretical limit of detection was estimated between 375 and 750 ppb, depending on the sensing material. A noteworthy conclusion from this sensitivity study is that the response of the chemical sensors scales almost linearly with the VOC concentration (at least for the 2 studied VOCs and in the studied concentration ranges). Finally, they have also confirmed a result which has already been highlighted with other technologies: extraction of kinetics parameters can help to increase classification performance.

Several improvements of the lab instrument are currently developed by the CREAB team and could, in the end, be transferred to Aryballe. First, an interesting approach to improve the sensitivity of the instrument is to vary the sensing temperature. J. Weerakkody et al. studied the effects of the temperature on the sensitivity of the chemical sensors, for temperatures ranging from 5°C to 45°C, and showed that a decrease in temperature leads to an increase of sensitivity [Wee+19; Wee+20]. This decrease in temperature can then help to considerably decrease the limit of detection of a VOC. This result could be of great relevance for applications under controlled conditions where a VOC has to be detected at really low concentration, such as biomedical applications. Second, to improve the diversity of the instrument (and consequently its overall performance, both in classification and quantification), new sensing materials need to be developed. Two ways are explored, namely Odorant Binding Proteins[15] and DNA, and have already proved their worth for other technologies. C. Hurot et al. investigated the use of OBPs as potential candidates for new sensing materials with an instrument operating in liquid phase [Hur+19]. They tested one OBP of a rat, designed 3 derivatives and immobilized them over the gold surface. They showed a good sensitivity and selectivity towards 3 different VOCs. However, authors stressed a limited lifespan of the sensing materials and a narrow linear range regarding the effects of the concentration, so further developments are required. In the same vein, S. Gaggioti et al. investigated the use of DNA as new sensing materials, in particular hairpin DNA (hairpin term comes from its hairpin-shaped structure) [Gag+20]. By using a prism combining peptides and hairpin DNA, they demonstrated similar results as the ones of Brenet et al., namely a good sensitivity and a good discrimination power. The studies of C. Hurot et al. [Hur+19] and S. Gaggioti et

[15] Odorant Binding Proteins (OBPs) play a key role in natural olfaction. In fact, odorant molecules do not directly go to the olfactory receptors by themselves but are transported by OBPs. Odorant molecules bind to these OBPs which transport them to the olfactory receptor then interact, or not, to activate, or not, the ORN. So OBPs are the first step of natural olfaction [Arc18].

al. [Gag+20] pave the way for an increase in the number of sensing materials boarded by Aryballe's instrument in the future.

The last point, about the addition of sensing materials, highlights a noteworthy advantage of Aryballe's eNose, and more widely of optical methods, compared to other existing technologies. Optical methods offer the possibility of increasing the diversity without increasing the overall complexity of the instrument. For instance, when the interactions are measured by a simple camera as in the case of SPRi or colorimetric sensors, the addition of sensing materials does not require the adjustement of the architecture of the system (*e.g.* eNose of Suslick's group can board up to 120 colorimetric sensors). That is a huge advantage over other transduction methods such as conductivity-based or mass-based methods. In fact, it can also explain why some authors in the field are so timorous regarding the increase in the number of chemical sensors: for a lot of well-established technologies, such as MOS-based sytems, increasing the number of sensors is costly. As we have already mentioned in Section 1.2.6, we believe that a key step for improving an eNose system is to increase the number of sensors, so it can presumably be used for a larger range of applications than a smaller sensor array.

1.4 Challenges of the thesis

As we have just seen, several results have been already reported regarding the performance of the instrument. Despite the significance of these results, they have been demonstrated with a lab instrument under lab conditions (for instance with dry air) with a lab setup, and consequently, with small data sets. Unfortunately, in the eNose field, it often happens that, what it is proven in the lab, never goes beyond the lab [Mar14]. This is especially true when proof-of-concept lacks realistic conditions, statistical reliability and external validation.

The miniaturization of the device by Aryballe is a key ingredient for alleviating all these issues. In fact, it allows the duration of the experiments to be greatly reduced. Let us do a quick comparison. In the study of Brenet et al. [Bre+18], one measurement (so one sample) with the lab instrument lasts for 30 minutes, which is common for such experiments. In this thesis, the total duration of an experiment will be max 6 minutes in Chapter 5 (about stereoselectivity), but will also be as short as a few seconds in Chapter 6 (about robot olfaction). This short experimental time is a chance to provide statistically reliable results, following the recommendations about the data set size that we have introduced in Section 1.2.6. The dimensions of the instrument are also a chance to get it out of the lab.

This thesis reports the first results obtained with Aryballe's eNose and several challenges, some of them discussed in Section 1.2.5, have been tackled. These results are quite complementary with the previous results found by the CREAB team [Bre18; Hur+19; Wee+19; Wee+20].

Classification performance (Chapters 3, 4, 5 & 6). Classification is one of the two main goals of an eNose, in combination with quantification. In fact, classification is always the first step, since it is a prerequisite for VOC quantification. A lot of issues are related to classification and it is much more than which classifier is the best to use (which is a difficult question

by the way). For instance, in this thesis, we will investigate feature extraction, focusing on features extracted either from the stationary or transitory regime (Chapter 3). Classification performance depends obviously on classes and how difficult the separation of the classes is. In molecular recognition, one of the hardest problems is the discrimination of two enantiomers of a chiral molecule [Tro14]. Two enantiomers are almost the same molecule and share almost the same physico-chemical properties, they are like the left and right hands but at a molecular level. In Chapter 5, we show that Aryballe's eNose is actually capable of such a discrimination power.

Unmixing or **Source Separation** (Chapters 3 & 6). In the field, VOCs are rarely, if not never, present as pure VOCs [Bou+03]. For an eNose, it is problematic due to the non-specificity of the chemical sensors. In fact, each chemical sensor can likely interact with all the components of the mixture. So, a mixture of VOCs will generate a signature just like a pure VOC but, a signature related in a non trivial way to the individual signature of each VOC and to their concentration. In Chapter 3, we define a theoretical mixture model which is non-linear and based on the Langmuir model. Under low concentration regime or low affinity regime, this model can, however, be linearized. This linearization is used in Chapter 6 for unmixing signals in a real-life application. We show that, despite the simplicity of the linear model, it can fairly estimate which VOCs are present with which intensities (however, we do not prove that these intensities relate directly to VOC concentration).

Drift (Chapter 4). Chemical sensors are prone to drift over time. Drift is responsible for the lack of stability of chemical sensors and is considered one of the main factors explaining the low number of practical applications with these sensors [Pad+10]. Concretely, it is very likely that the generalization of the classifier to a data base acquired few days or few weeks after the training set will be quite bad, due to the drift. In fact, it is still largely unrealizable to make sensing materials which do not drift over time [Ver+12]. So, the only way to compensate for the drift is to design post-processing methods to align as best as possible data distributions from the training and testing sets (the drift issue can be seen as a special case of "Transfer Learning" in Machine Learning). In Chapter 4, we discuss drift in detail and the main methods related to drift compensation. A class of methods has received little attention in the literature, namely methods which assume that there is no label available in the testing set. By considering this scenario, we propose two correction methods based on the idea that the drift between the two data sets (training and test) follows a preferred direction. During the thesis, we have acquired a comprehensive data set over 9 months which is affected by drift. We test several methods to compensate for the drift and show that the two proposed methods can remove a large part of the drift, enabling the classifier to better generalize.

Complex environment (Chapters 2 & 6). For a part of applications, eNoses are meant to be used in the field while studies are often carried out under controlled conditions in a lab environment [Tri10]. As an example, eNose could be used for monitoring environmental pollution [Bou+03]. In this case, it will continuously measure VOCs which are released in the environment. To name a few issues, environmental conditions can then no longer be

controlled, the VOC onset is then no longer known and needs to be identified, the subtraction of the baseline (reference VOC) is in turn more tricky, and VOCs can easily mix at variable concentrations. For monitoring environmental pollution, eNose can be static, waiting for VOCs to reach by themselves the instrument. But, other applications, such as gas leak detection, require in addition that the instrument can move and locate in space the VOC source [IWM12]. In the eNose literature, this field of applications is referred to as robot olfaction. In Chapter 2, we describe two robotic platforms which have been built during the thesis to test Aryballe's eNose in robot applications. In Chapter 6, we present several data sets from these two setups and tackle two main issues in robot olfaction: classification and unmixing.

To conclude, we would like to stress a current limitation of the results presented in this thesis. All the results are classification results, leaving aside quantification. In Chapter 6, we report some VOC intensities but, the true VOC concentrations are never measured, so a relation between these intensities and the VOC concentrations cannot be directly established yet.

Experimental setups

2

Contents

2.1 Introduction

During this thesis, several experimental setups have been designed for the generation of several data sets aiming at studying various issues. The introduction of these setups and some preprocessing steps are gathered in this chapter. However, the introduction of the data sets are detailed in the involved chapters.

- **Automatic valve** (Section 2.2). This setup has been built by Aryballe Technologies and aims at generating large data sets under controlled conditions. This setup has been especially useful for showing that the instrument is "stereoselective", in Chapter 5.
- **Sniffer robot** (Section 2.3.1). This setup has been built and designed at Gipsa-Lab and aims at studying the performance of the device in a robot scenario. This setup has been especially useful for both the classification of 12 different VOCs disseminated in an uncontrolled environment (Chapter 6) and the study of the drift over 9 months (Chapter 4).
- **Sniffer arm** (Section 2.3.2). This setup has been built and designed at Gipsa-Lab and aims at studying the performance of the device in a robot scenario in which gas mixtures occur. This setup has been especially useful for the study of gas mixtures under uncontrolled conditions (Chapter 6).

The generation of the data sets and the building of the setups have been assisted by several persons. I especially would like to acknowledge, for their highly valuable help,

- **Romain Dubreuil**, **Johanna Decorps** and **Etienne Bultel**, who are working for Aryballe Technologies and helped me many times to carry out the experiments with the automatic valve (introduced in Section 2.2).
- **Leonardo Carlos Dolcinotti**, who was intern at Gipsa-Lab during 3 months under the supervision of Simon and myself. He built a line-follower robot (introduced in Section 2.3.1).
- **Xiran Zuo**, who was intern at Gipsa-Lab during 3 months under the supervision of Simon and myself. She especially helped me with the first experiments using the Sniffer arm (introduced in Section 2.3.2).
- **Sophie Min**, **Rémy Jaccaz**, **Aurélien Carriquiry**, **Sylvain Géranton**, **Anton Andreev**, who are working at Gipsa-Lab ("*Plateformes service*") and who substantially helped to build the robotic setups (introduced in Section 2.3.1 and Section 2.3.2).

2.2 Automatic valve

Goal. This setup aims at generating data using a method called 3-phase sampling. This sampling method has been detailed in Introduction (Section 1.2.2) and relies on 3 steps: baseline acquisition, VOC injection and recovery. The setup is fully automatic in order to generate large data sets ($\geqslant$100 samples/class).

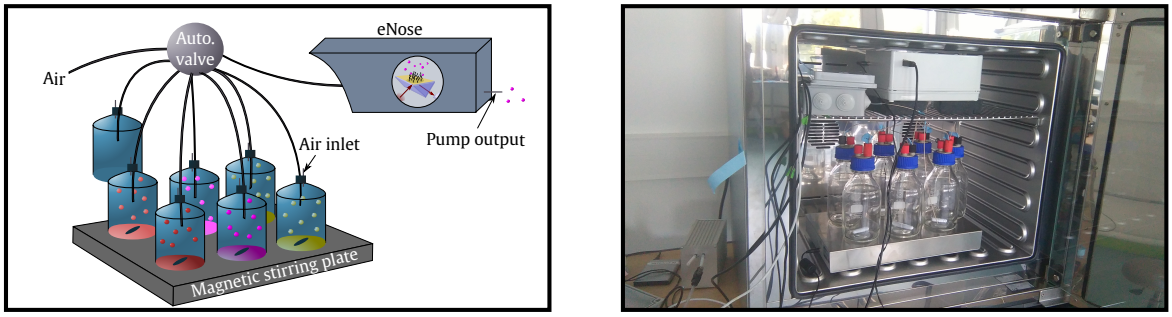


Fig. 2.1.: Automatic valve. Left figure is a graphical representation of the setup and right figure corresponds to reality (the setup in placed in an oven).

Description. The main part of this system is an automatic valve represented in Figure 2.1. In total, 8 lines can be allocated. But, one of the 8 lines is booked for ambient air which is used as our reference gas during all the experiments. Except the reference line, the lines are inserted in vials (whose the capacity depends on the experiment), which are sealed with a Teflon-coated silicon septum and filled with a liquid solution of each VOC studied. The vials containing the liquid solutions are agitated using a magnetic stir bar to homogenize the headspace and a small PolyEtherEtherKetone (PEEK) tube is also used to balance the pressure inside the vials. The fluidic system is made of PEEK tubing. Each gas line has the same tube length to prevent any "artificial" discrimination due to the transport to the instrument.

The valve is programmed to start with the air line (baseline acquisition). Then, at t_s, the valve switches to the line i (VOC injection). At t_e, it goes back to the air line (recovery). The process is repeated for the line $i + 1$ and so on.

In this setup, temperature can be controlled by placing the entire setup in an oven. But it can also be placed out of the oven in order to generate more realistic data.

Baseline manipulation. A basic preprocessing with this kind of data is the baseline subtraction. One reason for baseline subtraction is that baseline tends to shift over time. In Figure 2.2, we represent four different measurements of a single chemical sensor, taken at different times. We note a clear drift of the baseline. Its subtraction helps to increase the repeatability of the instrument. This step is straightforward in this setup since we know exactly when the VOC is injected. Taking as example Figure 2.2, we simply take the average of the 25 first seconds of signal for each chemical sensor and subtract this value from the whole response. We therefore no more have Reflectivity (%) but ΔReflectivity (%) values. For the next setups, we will see that the baseline subtraction is not as straightforward.

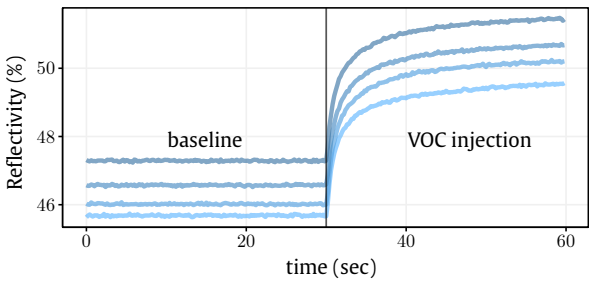


Fig. 2.2.: Four measurements of a single chemical sensor over time. A baseline drift is clear.

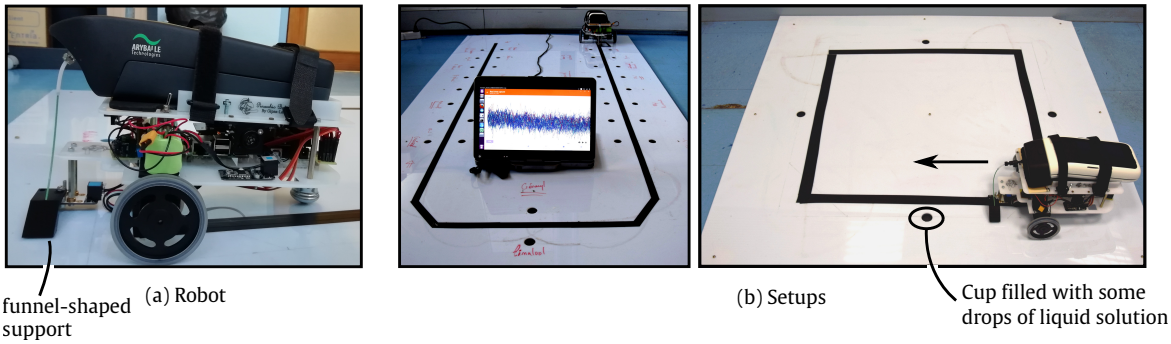


Fig. 2.3.: Sniffer robot platform. (a) Line follower robot used. (b) The two setups built, only the size and the number of possible gas sources differ from one setup to another (up to 34 gas sources for the setup to the left and up to 4 gas sources for the setup to the right).

2.3 Robotic platforms

Two different robotic setups have been designed during the thesis, namely a line-follower robot and a robotic arm. Their goal is to evaluate Aryballe's eNose in a complex environment.

2.3.1 Sniffer robot

Goal. Sniffer robot aims at evaluating gas recognition in this thesis, in a robot scenario. The platform consists of a robot (see Figure 2.3a) that carries the eNose, moving over a flat surface where gas sources are placed (Figure 2.3b). Some data sets from this setup will be introduced in Chapters 4 & 6.

Description. The robot is a line-follower robot, meaning that it can follow a black line drawn over a white surface. We use two optical sensors to achieve this function. Currently, the robot is neither autonomous nor capable of locating itself in space. Besides, the robot is not yet capable of interacting with the eNose, so the measure of the eNose and the movement of the robot are independent. In fact, we are interested here in testing the eNose in a practical application (robot olfaction) rather than in defining an optimal movement strategy. The speed of the robot is tuned before the experiment and is generally set between to 2 cm/s and 3 cm/s. To generate substantial data sets, the path is followed multiple times by the robot.

A funnel-shaped support has been made with a 3D printer and the injection tube of the eNose is inserted in this support in order to increase the suction area. The ground is a polycrystal plate which is lifted by 1.5cm. The dimensions of the ground is either 1m $\times$ 1m $\times$ 2.5mm for the small setup (up to 4 gas sources, used in Chapter 4) or 2m $\times$ 2m $\times$ 2.5mm for the larger one (up to 34 gas sources, used in Chapter 6).

Gas sources are small cups which are slid below the plate where the holes have been made. These cups are filled with some drops ($\sim$250μL) of liquid solutions of studied VOCs, just before the experiment. These setups are placed in an indoor environment, basically a normal office with low natural advection (no ventilation system). Humidity and temperature are recorded but not controlled. Regarding parameters of the eNose, the frame rate of the camera is always set to 5 Hz and the airflow usually to 60 mL/min. The frame rate can be increased

and the airflow can be decreased but these values are sufficient in practice to measure the chemical reactions and their dynamics.

Data. In Figure 2.4, we represent the response of one chemical sensor during one lap. We can note that the data look like peaks and contrast a lot with the data generated according to a 3-phase sampling (see Figure 2.2 for comparison). This is due to the very short time injection (∼1 second compared to 30 seconds for the measurements in Figure 2.2).

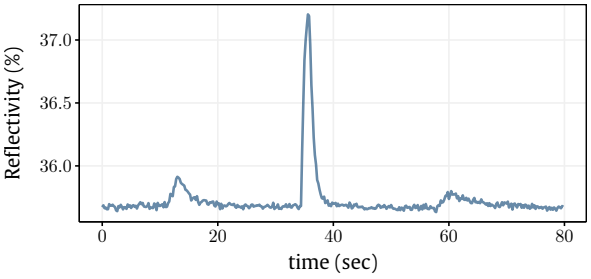


Fig. 2.4.: Data obtained with the Sniffer robot. We represent the response of one chemical sensor during one lap with the small setup. Each peak corresponds to the passage of the robot over a cup.

Video. The reader is encouraged to watch a video[1] that I have made during the thesis using the large setup (Figure 2.3b left):

http://simonbarthelme.eu/personal-website/videos/robot12VOC.mp4

2.3.2 Sniffer arm

Goal. Sniffer arm aims at generating gas mixtures in this thesis, in a robot scenario. For instance, gas mixtures can easily occur in the event of two gas sources which are spatially close.

Description. The main part is an aluminium trapezoidal shaft that moves the eNose along one dimension (see Figure 2.5). The total length of this path is 36.5cm. Again, a funnel-shaped support is added to increase suction area. Gas sources are scent strips, directly placed along the 1D-path, on which few drops are deposited. The speed of arm movement is set to 1 cm/s.

In practice, several sweeps are performed in order to generate substantial data sets. At the end of each sweep, the arm stays in place for 20 sec. The response of the eNose during the sweeps is recorded continuously (without any interruption between each sweep). Again, Sniffer arm and the eNose are independent and cannot interact yet.

Data. In Figure 2.6, we represent the response of one chemical sensor for two different scenarios that one can encounter in the field. The whole data set, another scenario and an unmixing algorithm will be introduced later in this thesis, in Chapter 6.

[1] If the link does not work, try this one:
https://drive.google.com/file/d/1AddmCYFwQcHtG1aTksn8zyC-0e7CDzbU/view

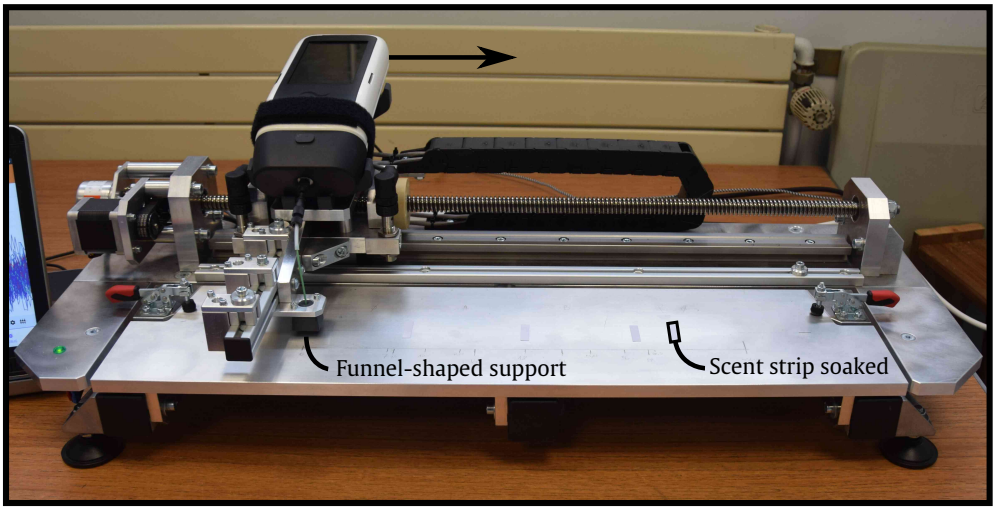


Fig. 2.5.: Sniffer arm.

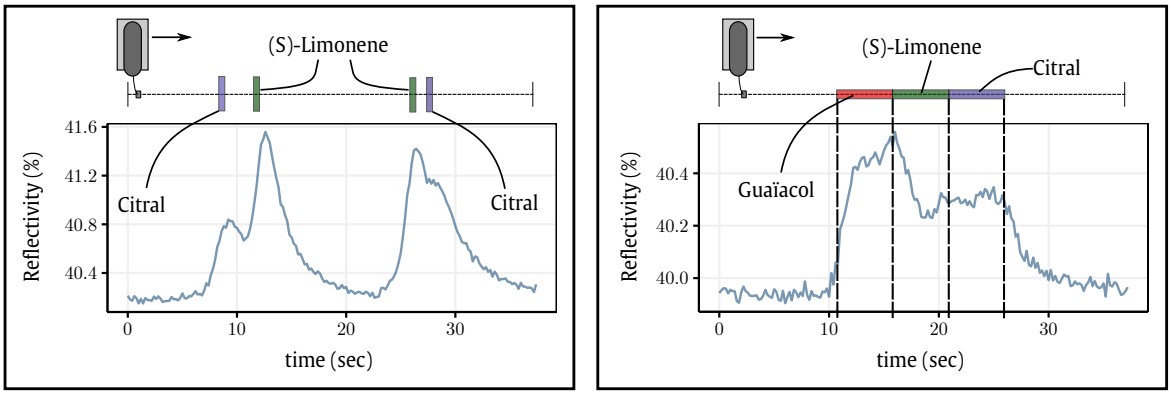


Fig. 2.6.: Data obtained with the Sniffer arm. We represent the response of one chemical for two different scenarios which will be described in Chapter 6. Left figure: two gas sources are spatially close along the path. Right figure: the arm goes over successive trails of pure compounds.

Video. The reader is encouraged to watch a video[2] that I have made during the thesis using this setup:

http://simonbarthelme.eu/personal-website/videos/closeSources.mp4

2.3.3 Preprocessing issues

Baseline manipulation. Compared to the automatic valve, baseline manipulation is more tricky here. Indeed, in Section 2.2, the injection time is exactly known so the baseline is perfectly identified in the signal. This makes the first preprocessing step easy but this is no longer the case with the robotic platforms and we do not know exactly when VOCs are injected.

A trivial solution would be to take the first measurement points to estimate the baseline (assuming that the robot does not start around a gas source). However, baseline drifts over

[2]If the link does not work, try this one:
https://drive.google.com/file/d/14exr-R1c4hTyitn1WeuqKYA3uLgDrI5z/view

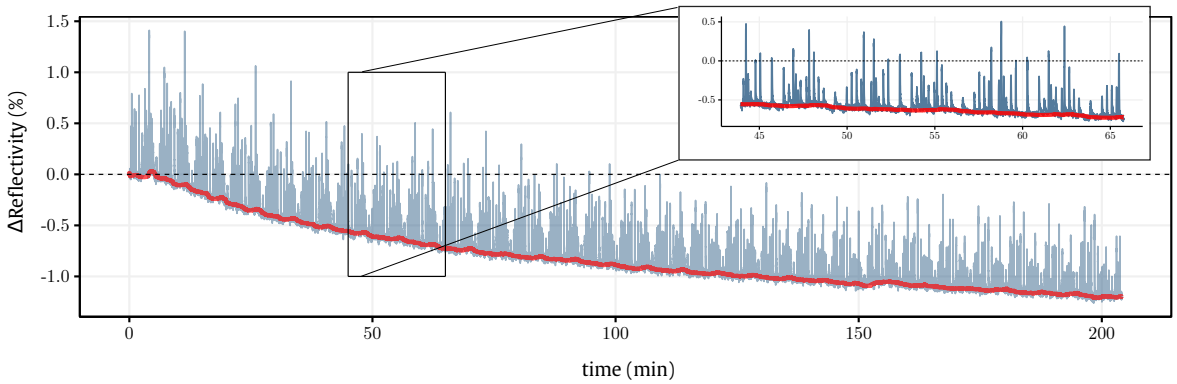


Fig. 2.7.: Baseline drifts over time. Drift estimation is represented in red.

time due to complicated phenomena which take place during an experiment. For instance, reference gas (ambient air) is varying over time due to the evaporation of the studied VOCs. Temperature and humidity can also vary. In Figure 2.7, we report the response of one chemical sensor during more than 3 hours using the Sniffer robot. Baseline drift is clearly perceptible and may be attributed to variation in temperature due to the heating of the electronic system.

If the first points cannot be used, then baseline drift must be estimated over time. The difficulty is that VOC injections and baseline drift appear at the same time, even if VOC injections are much shorter over time. Since baseline drift is a slow-varying process compared to injection areas, we could use some frequency methods and remove low frequencies. Here, we prefer to directly estimate the baseline trend.

We note $y_p(t)$ the response of the chemical sensor p at time t. The approach proposed is based on quantile filtering which enables to estimate the trend by avoiding the peaks. By considering an integer k and a scalar $q \in [0, 1]$, the estimation of the baseline corresponds to the q-quantile of $\{y_p(t-k+1), ..., y_p(t)\}$. The estimation of the drift is reported in red in Figure 2.7.

The size k of the time window and the scalar q can be important, here they have been manually selected for the experiments. We set the parameters to $k = 100$ seconds and $q = 0.1$, for all the data generated with the robotic platforms.

Segmentation. Compared to the automatic valve, time series are not automatically segmented for robotic setups and peaks appeared at unknown times. So, a segmentation step is required to extract the peaks from the data (see Figure 2.8), and is detailed in Appendix A.

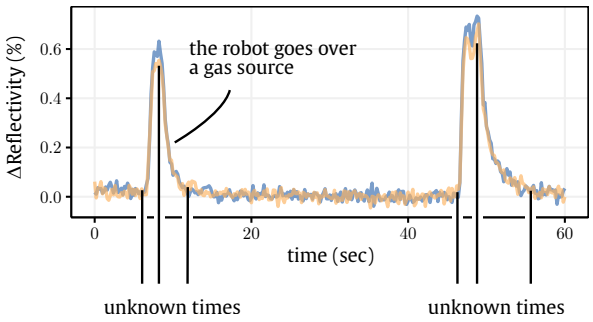


Fig. 2.8.: Time series of two chemical sensors using the Sniffer robot. In this figure, the robot went over two gas sources, corresponding to the two observed peaks.

Modeling

3

Contents

"After a few weeks Grenouille had mastered not only the names of all the odours in Baldini's laboratory, but he was also able to record the formulas for his perfumes on his own and, vice versa, to convert other people's formulas and instructions into perfumes and other scented products. And not merely that! Once he had learned to express his fragrant ideas in drops and drams, he no longer even needed the intermediate step of experimentation. [...] He had learned to extend the journey from his mental notion of a scent to the finished perfume by way of writing down the formula."

Patrick Süskind, *Perfume, the Story of a Murderer*[1].

In this extract, Grenouille is learning how to make a fragrance by the book, from the perfumer Baldini. Rapidly, the apprentice surpassed the master and Grenouille discovered how to go from the scent to the chemical formula, and *vice-versa*. In fact, we could say that he discovered the model which allows to go from the formulas of the molecules to the sensations that they bring.

Modeling is about the characterization of a system and a model tells us about the output that we should expect if we excite the system with a given input. In fact, if you know the model then, with a known input, we should be able to predict the output without carrying the experiment (like Grenouille). In our case, we will not go as far as Grenouille, tracing back to the structure of the molecules. Instead, we will try to describe more modestly the relation between the response of the instrument and two parameters: the VOC concentration and the affinity between the VOC and the chemical sensors.

The chapter is organized as follows. First, we detail a physico-chemical model for describing the time series in the presence of a pure VOC. It is the result of the combination of the Langmuir model, describing the chemical interactions, with a linear approximation of the response obtained by Surface Plasmon Resonance imaging. Then, the parameters of the proposed model is used to extract features from the dynamics, to test other features than the classical steady state. Finally, we formulate a model of the response during equilibrium phase when we no longer measure a pure VOC but a gas mixture.

3.1 Physico-chemical model

3.1.1 Langmuir model

Recall that a measurement of a chemical sensor results from the interaction between a VOC (we assume a pure VOC here), say A, and the sensing material, say S, of the chemical sensor. The chemical interaction is a **binding reaction** in the case of Aryballe's eNose, which can be written in the simplest case as:

$$\mathrm{A} + \mathrm{S} \underset{d}{\overset{a}{\rightleftharpoons}} \mathrm{AS} \tag{3.1}$$

[1]Translated from the German by John E. Woods.

Eq. (3.1) means that when the VOC A interacts with the sensing material S, it creates a "new" molecule, the complex AS. This is called **adsorption** of A on the surface covered by S. If the reaction was irreversible, it would go only from one side (A+S) to the other side (AS) and the complex AS could never become A+S again. Fortunately, with an eNose, reactions are always reversible, meaning that the complex AS can become A+S again. This is called **desorption** of A from the surface covered by S.

This reaction is controlled by two parameters which are naturally positive: the **adsorption constant** a (unit concentration^{-1}.time^{-1}) and the **desorption constant** d (time^{-1}). a and d are **specific** to the VOC A and to the sensing material S and are related to their **affinity**. a tells us the ability of A to bind easily to S, while d tells us the ability of A to leave easily S. For instance, if d is quite low, then the desorption can be quite long and so for the recovery phase (so we will have to wait for a long time before the response go back to the baseline value). These parameters a and d are sensitive to temperature, so we assume that the temperature is controlled during the experiment and does not change.

To describe mathematically the reaction in Eq. (3.1), we generally use a simple notion: the notion of free/occupied sites (see Figure 3.1). At the beginning of the experiment, the sensing material can be seen as a number of free sites on which a VOC can bind: this quantity of free sites is modeled by a surfacic concentration $[\mathrm{S}]_s$ (surfacic because the sensing material is fixed over a surface). When the VOC A is injected, A adsorbs on the sensing material and binds to some sites which become then occupied: the quantity of occupied sites is modeled by a surfacic concentration $[\mathrm{AS}]_s$. The number of free sites becoming occupied at time t can be quantified by a simple quantity: a (the adsorption constant) $\times[\mathrm{A}](t)$ (the concentration of A) $\times[\mathrm{S}]_s(t)$ (the number of free sites). Due to the reversibility of the reaction, the VOC A can desorb from the surface and the occupied sites can become free again. The release of A in the air can be quantified by a simple quantity: d (the desorption constant) $\times[\mathrm{AS}](t)$ (the number of occupied sites). This allows us to model the number of sites becoming occupied over time, reflecting the interaction:

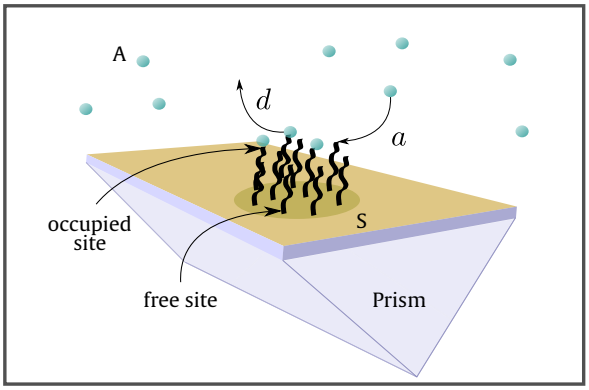


Fig. 3.1.: Free/occupied sites.

$$\begin{aligned}
\frac{\mathrm{d}[\mathrm{AS}]_s(t)}{\mathrm{d}t} &= \begin{array}{c}\text{number of free sites}\\ \text{becoming occupied}\end{array} - \begin{array}{c}\text{number of occupied sites}\\ \text{becoming free}\end{array} \\
&= a[\mathrm{A}](t)[\mathrm{S}]_s(t) - d[\mathrm{AS}]_s(t) \\
&= a[\mathrm{A}](t)\big([\mathrm{S}]_s - [\mathrm{AS}]_s(t)\big) - d[\mathrm{AS}]_s(t)
\end{aligned} \tag{3.2}$$

In fact, instead of quantifying the number of occupied sites, we sometimes prefer to model the **fraction of occupied sites**, defined as $\theta(t) = \frac{[\mathrm{AS}]_s(t)}{[\mathrm{S}]_s}$.

From Eq. (3.2), we can easily obtain a differential equation with respect to $\theta(t)$:

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = a[\mathrm{A}](t)\big(1-\theta(t)\big) - d\theta(t) \tag{3.3}$$

Eq. (3.3) is the simplest way to model over time the chemical interaction resulting from a VOC adsorption and comes from the seminal paper of I. Langmuir [Lan18]. It relies on some assumptions which could be violated in practice:

- the sensing surface is perfectly flat.
- the interaction is monolayer[2].
- all the sites are equivalent.
- there is no other interaction than the interaction between A and S (*e.g.* A does not interact with A).

Eq. (3.3) is a simple first-order linear differential equation in function to θ. Finding an analytical solution is easy only if $[\mathrm{A}](t)$ (the variations in concentration of A over time) is known and simple. In the context of 3-phase sampling, if air moves enough quickly in the pipes, then concentration follows approximately a gate function. In other words, the concentration of A is constant during all the injection, of value c, and is null outside the injection. Under this hypothesis, if we note t_s the beginning of the injection, the problem becomes:

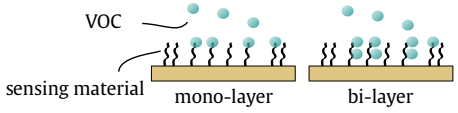
$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = ac\big(1-\theta(t)\big) - d\theta(t) \quad \text{s.t.} \quad \theta(t=t_s) = 0 \tag{3.4}$$

which is is straightforward to solve, if we note t_e the end of the injection (recall that $c = 0$ if $t > t_e$), then:

$$\theta(t) = \begin{cases} 0 & \text{if } t < t_s \\ \dfrac{ac}{d+ac}\big(1-e^{-(ac+d)(t-t_s)}\big) = \theta_a(t) & \text{if } t \in [t_s, t_e] \\ \theta_a(t_e)e^{-d(t-t_e)} & \text{if } t > t_e \end{cases} \tag{3.5}$$

In Figure 3.2, we represent the influence of each parameter on the global response. First, it is interesting to note that the adsorption constant a and the VOC concentration c play symmetrical roles in the model (Eq. (3.5)). The greater these two parameters are, the faster the adsorption is and the greater the response is. However, they do not affect the desorption. In contrast, the desorption constant d affects both the adsorption and the desorption. The greater d is, the faster the adsorption and the desorption are but the lower the response is. Second, θ is the fraction of occupied sites, so it is a value naturally bounded by 0 and 1. Finally, we can notice that if we inject A for a sufficiently long time, the response of the model stabilizes around a value ($\frac{ac}{d+ac}$): this value corresponds to the steady state value and means

[2] Monolayer assumption means that each site can hold only one molecule of A. Multi-layer assumption (e.g. bi-layer) would assume that each site can contain multiple molecules of A.



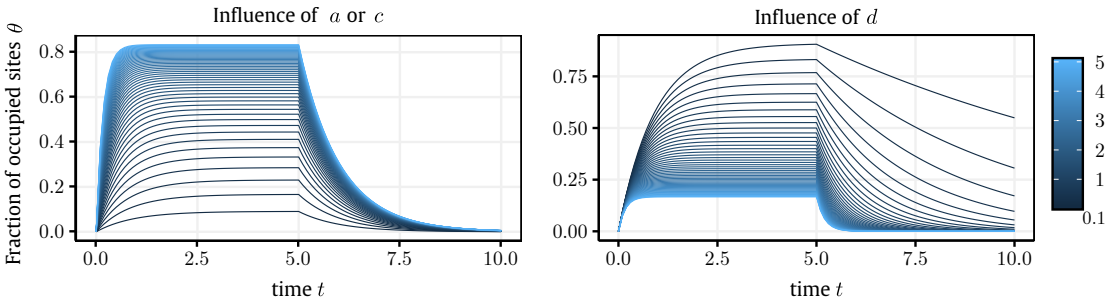


Fig. 3.2.: Influence of the parameters on the response. When one parameter varies (from 0.1 to 5), the others are fixed to 1. a and c have symmetrical roles in the model so their influence is the same.

that the chemical reaction is at the equilibrium phase[3]. Steady state value is classically used as a feature for classification.

This model is called the **Langmuir model**[4] in reference to Irving Langmuir [HBZ06; Lan18]. This model describes only the chemical interaction, but the response of Aryballe's eNose results from some physical phenomena that we have to model too.

3.1.2 SPRi model

The transduction method used here is based on **Surface Plasmon Resonance imaging** (SPRi). It is a bit technical so we provide here only a short introduction to SPRi, but readers can find more details in the book of J. Homola [Hom06]. We also recommend the reading of the thesis of E. Maillart [Mai04] for French speakers.

SPRi is actually the combination of two physical phenomena, namely two electromagnetic waves which exist under certain conditions. The first one is the **surface plasmon** (SP) which is an electromagnetic wave existing at any metal-dielectric interface (*e.g.* the interface gold-glass). The second one is an **evanescent wave** created by an incident light which is totally reflected by a surface (meaning that the incidence angle is greater than the critical angle). This evanescent wave appears from both sides of the surface and decays exponentially in the direction perpendicular to the surface (it is thus limited to a given depth $L_z \approx 100$ nm). In fact, that is why this wave is called evanescent: it disappears quickly. This evanescent behaviour is quite interesting because it will not be sensitive to what occurs beyond few hundreds of nanometers from the surface: it will be sensitive to phenomena which occur only quite close to the surface (in our case, binding reactions) [Mai04]. The conditions necessary to the creation of these two waves can be obtained by using a prism coated with gold (the metal-dielectric interface) and a monochromatic light (for creating the total reflection). This

[3] The chemical reaction is always reversible, enabling the adsorption and desorption of A. When the response reaches the steady state value, it means that these two events became sort of equiprobable, causing the response to no longer change: this is the equilibrium.

[4] We can notice that this model is very similar to another well-known model: the model describing the charge and discharge of a capacitor. However, the Langmuir model allows us to understand the connections between the model parameters and the chemical parameters of the reaction (in particular it tells us the influence of the VOC concentration on each part of the response).

system is represented in Figure 3.4, in Kretschmann configuration (the configuration used by Aryballe's eNose).

When the evanescent field is created by a TM-polarized (TM: Transverse Magnetic) incident light, it can be used to excite the surface plasmons in a resonant manner. By exciting the surface plasmons, a part of the energy of the incident light is transferred to the plasmons, so the amount of reflected light decreases due to the coupling. Here, the coupling of these two waves is performed by a change of the incidence angle. The decrease in the amount of reflected light depends on this angle and can be caught by a simple camera (the detector). By playing with this incidence angle, we can trace the reflectivity variations as a function of the angle, leading to the so called **plasmon curve** represented in Figure 3.3. The reflectivity is minimal for a given angle, namely the SPR angle. During a classical experiment with SPRi, we start by generating the plasmon curve and the angle is then fixed to a given value, which is not the SPR angle. This working angle, Θ_{opt}, is preferably located in the linear range of the plasmon curve and no longer changes during the rest of the experiment. For Aryballe's eNose, we do not plot the plasmon curve before each experiment (it is time-consuming). Aryballe defined reflectivity ranges which correspond to the linear range of the plasmon curve. In practice, the working angle is fixed such that the reflectivity measured is around 60%.

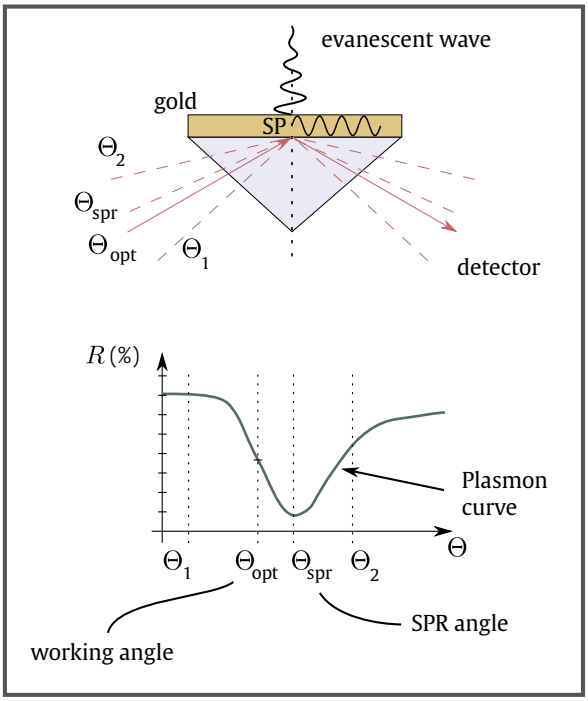


Fig. 3.3.: Working principle of SPRi and plasmon curve generation. By varying the incidence angle Θ, the evanescent wave and the surface plasmons (SP) can be coupled. For a given range of Θ, a part of the energy of the incident light is transferred to the SP. This loss in energy is perceived as a loss in reflectivity. The plasmon curve shows the evolution of the reflectivity according to Θ. The angle for which the curve is minimal is the SPR angle Θ_{spr}. The working angle Θ_{opt} is preferably located in the linear range of the curve. See the text for more details.

Why use SPRi to monitor binding reactions? In fact, the entire plasmon curve is quite sensitive to the medium near the metal interface. Indeed, any tiny increase in the refractive index of this layer will result in a simple shift to the right of the plasmon curve. Thus, when a VOC binds to the sensing material which is immobilized on the golden surface, the plasmon curve is shifted (to the right). Figure 3.4 shows the shift of the plasmon curve during an experiment. This shift results in an increase of reflectivity at the fixed angle Θ_{opt}. So, before injection (baseline), the measure is at a certain reflectivity which then increases after VOC injection. A simple difference between the two values leads to our raw recording, namely

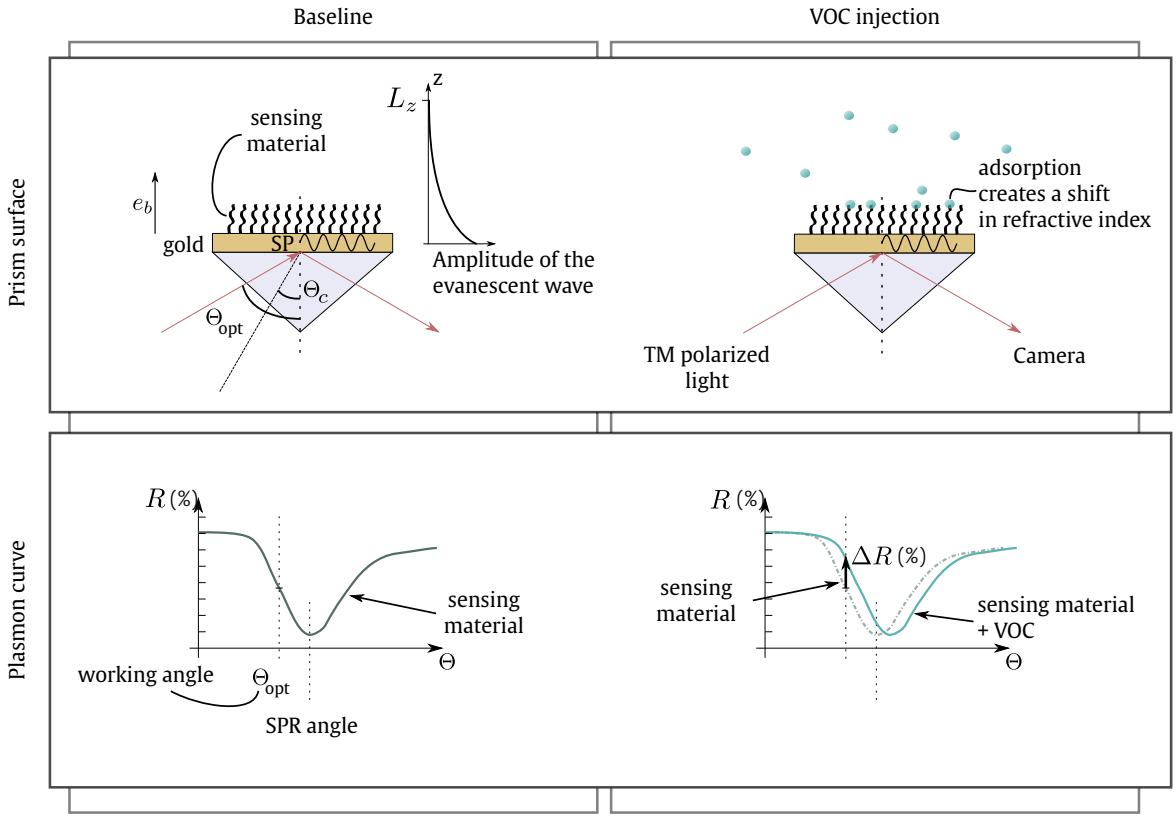


Fig. 3.4.: Working principle of SPRi. When adsorption occurs, it changes the refractive index, which in turn shifts the plasmon curve. See the text for more details. Surface plasmons are abbreviated as SP.

a delta-reflectivity ΔR (%). In our case, each pixel of the images from the camera gives a reflectivity measurement representing the interaction on a spatial area of the prism.

However, something is missing in our explanation. In fact, a simple camera does not return automatically reflectivity measurements but pixel values (in our case, 8-bit pixels, so a value between 0 and 255 telling us about the amount of light reflected by the surface). How can we convert pixel values into **reflectivity** measurements? In fact, the phenomemon occurs only when the incident light is TM-polarized and completely disappears if the light is TE-polarized (TE: Transverse Electric). In other words, with TE-light, there is no excitation of the surface plasmons, so the incident light does not lose any energy, and so the amount of reflected light is a good approximation of the maximum light that we can receive from the surface [Bas+03]. Consequently, we can see the image formed with a TE-polarized light as a good reference of max light and a simple division of the TM-image by the TE-image gives us a reflectivity image. In Figure 3.5, we show some images obtained from real experiments.

A last question is then, how to relate ΔR to the number of sites becoming occupied over time (recall that it is the quantity characterized by the Langmuir model)? The answer could be found by using Maxwell's equations, which leads to a complex model [Mai04]. However, this model can be well approximated in liquid phase (and we assume that this model is still true

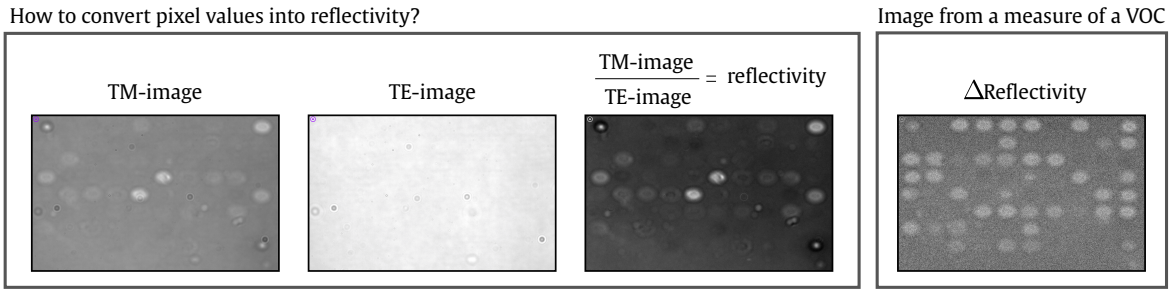


Fig. 3.5.: Conversion of pixel values into reflectivity. On TM-image, each light area corresponds to sensing materials on the surface. On TE-image, we can no longer see sensing materials since the excitation of surface plasmons can no longer occur. The division of TM-image by TE-image (multiplied by a factor depending on the refractive index of the prism) gives us a reflectivity image. Right image corresponds to a ΔReflectivity image obtained by subtracting an image from a VOC injection by an image from the baseline. The adsorption of the VOC is clear, since the light areas remain after subtraction, indicating that there is more material on the surface due to the bindings.

in gas phase) by a linear model, when the thickness e_b (see Figure 3.4) of the medium is much smaller than the depth of the evanescent wave L_z [Ste+91; Jun+98]:

$$\Delta R(t) = \frac{S}{L_z}\frac{\partial n}{\partial c} \times \Gamma(t) \tag{3.6}$$

where $S = \frac{\partial R}{\partial n}$ is the SPRi sensitivity (the variation of reflectivity induced by a variation of refractive index), $\frac{\partial n}{\partial c}$ (cm^3/g) is the change in refractive index of the VOC with change in concentration and finally Γ (ng/mm^2) is the recovering rate of the VOC. This model is reasonable for two reasons: the thickness of the sensing material is quite small (∼2-3 nm), and VOCs are generally small molecules (∼1-2 nm, otherwise they could not evaporate easily). So, the adsorption of the VOC on the sensing material will result in a medium of max 5 nm which is much smaller than L_z (≈ 100 nm). To bring the fraction of occupied sites θ out, we rewrite Γ as $\Gamma = [\mathrm{S}]_s m\theta$ [Hom06] (p78) where m is the molecular mass of the VOC.

$$\Delta R(t) = \frac{S}{L_z}\frac{\partial n}{\partial c}[\mathrm{S}]_s m \times \theta(t) \tag{3.7}$$

3.1.3 Full model

The SPRi model tells us that the SPRi response is simply proportional to the chemical response. By merging the SPRi model with the Langmuir model, we can build a global model of the measurement $y(t)$ over time:

$$y(t) = \begin{cases} 0 & \text{if } t < t_s \\ \dfrac{S}{L_z}\dfrac{\partial n}{\partial c}[\mathrm{S}]_s m \times \dfrac{ac}{d+ac}\left(1 - e^{-(ac+d)(t-t_s)}\right) = y_a(t) & \text{if } t \in [t_s, t_e] \\ y_a(t_e)e^{-d(t-t_e)} & \text{if } t > t_e \end{cases} \tag{3.8}$$

We can easily see that the model (3.8) depends on many different parameters. To see the relevance of each parameter, let us highlight the dependences of the parameters to the dimensions of our problem (we have several chemical sensors and, we study several VOCs).

For that, we note p and P one chemical sensor and the total number of chemical sensors. We note r and R one VOC and the total number of VOCs studied.

For the SPRi model, the SPRi sensitivity S and the depth penetration L_z are dependent on SPRi settings (*e.g.* wavelength, metal-dielectric interface, ...), so they are neither dependent on the VOC nor the chemical sensor [Hom06]. However, $\frac{\partial n}{\partial c}$ is the amount of change of refractive index per VOC concentration unit so it depends on the VOC r, and so for the mass m. $[\mathrm{S}]_s$ corresponds to the number of free sites so it clearly depends on the chemical sensor p. $\theta(t)$ is the fraction of occupied sites on the sensing material by the VOC, so it depends both on p and r. By stressing the problem dimensions, we can rewrite Eq (3.7) as:

$$\Delta R_{pr}(t) = \alpha_{\text{spr}} \times \eta_r \times \phi_p \times \theta_{pr}(t) \tag{3.9}$$

with $\alpha_{\text{spr}} = \frac{S}{L_z}$ the parameters depending on SPRi, $\eta_r = \frac{\partial n}{\partial c} m$ the parameters depending on the VOC and $\phi_p = [\mathrm{S}]_s$ the parameter depending on the chemical sensor.

For the Langmuir model, the adsorption constant a and desorption constant d depends on the chemical reaction, so both on p and r. The VOC concentration c depends only on r.

The full model becomes:

$$\boxed{y_{pr}(t) = \begin{cases} 0 & \text{if } t < t_s \\ \nu_{pr} \times \dfrac{a_{pr} c_r}{d_{pr} + a_{pr} c_r} \left(1 - e^{-(a_{pr} c_r + d_{pr})(t - t_s)}\right) = y_{a,pr}(t) & \text{if } t \in [t_s, t_e] \\ y_{a,pr}(t_e) e^{-d_{pr}(t - t_e)} & \text{if } t > t_e \end{cases}} \tag{3.10}$$

with $\nu_{pr} = \alpha_{\text{spr}} \eta_r \phi_p$.

3.1.4 Identifiability

For fitting the model (3.10) to some real measurement $\boldsymbol{y}_{pr} \in \mathbb{R}^{N_t}$ of duration N_t, we need to find the set of parameters $\boldsymbol{\zeta}_{pr} = (\nu_{pr}, c_r, a_{pr}, d_{pr})$. For that, we need to define a cost function which measures the amount of errors made by the model, and solve an optimization problem. A classical way is to consider the least squares cost function:

$$\hat{\boldsymbol{\zeta}}_{pr} = \underset{\boldsymbol{\zeta} \geqslant 0}{\arg\min} \left\| \boldsymbol{y}_{pr} - f(\boldsymbol{\zeta}) \right\|_2^2 \tag{3.11}$$

with f the function described by the model (3.10) for which (t_s, t_e) are two known constants. f returns a vector of dimension N_t.

The minimization problem in Eq. (3.11) implies a preliminary study: **identifiability** of the model.

A model is said identifiable if the parameters of the model can be identified uniquely. In our case, it would mean that there exists a unique set of parameters $\boldsymbol{\zeta}^*_{pr} = (\nu^*_{pr}, c^*_r, a^*_{pr}, d^*_{pr})$ for which the cost function is minimal. Identifiability is quite important if the parameters

need to interpretable. For instance, interpretability is required for the VOC concentration c_r. Identifiability is also required for using the parameters as new features for classification. Let us illustrate this point. Assume that there are two different solutions ζ_1 and ζ_2 which have the same cost and we cannot tell which one is the best. So, the algorithm solving the optimization problem (3.11) will sometimes converge to ζ_1 and will sometimes converge to ζ_2, depending on the initialization. Since initialization is often either random or arbitrary, this will cause a problem for Machine Learning algorithms which assume that the signature of a VOC is kind of "unique".

In the case of the optimization model (3.11), it is easy to see that the parameters a_{pr} (adsorption constant) and c_r (VOC concentration) are not identifiable. In fact, we say that the model suffers from an **indetermination**. For any $\lambda \in \mathbb{R}^*$, the solutions $\frac{a_{pr}}{\lambda}$ and $c_r\lambda$ have the same cost. A solution is to rewrite the model (3.10) by replacing $a_{pr}c_r$ by a single parameter z_{pr}. However, even with this simplification, adsorption part suffers from another indetermination.

The rest of the identifiability analysis is left in Appendix B. In fact, we will see in the next section that the model in its current form suffers from a main problem: the constants in the exponential terms are not well predicted by the Langmuir model.

3.2 Experimental

3.2.1 Data sets

To validate or invalidate the model (3.10), we have generated 2 different data sets from two different eNoses using two different ways of acquisition.

One data set is generated using an automatic valve described in Chapter 2. The other data set is generated by hand: the vial containing the VOC is brought to the instrument by a human operator. Both data sets were generated using a 3-phase sampling from the headspace of a vial: baseline acquisition, VOC injection, recovery. In Table 3.1, we report some information about the two data sets.

3.2.2 Model validation

Model validation. In the previous section, we saw that the parameters a_{pr} and c_r are not identifiable and need to be replaced by a common parameter $z_{pr} = a_{pr}c_r$. To estimate the parameters $(\nu_{pr}, z_{pr}, d_{pr})$, we use a quasi-Newton algorithm called L-BFGS [Byr+95] with the `R` package `optim`. All the parameters are constrained to be non-negative.

The fit of the model (3.10) to real data performs in fact poorly regarding the estimation of the parameters: z_{pr} is close to 0 and the estimated parameters depend a lot on the initialization.

Name	Acquisition	VOCs	Time segmentation		f_s	P (M)	N
Data set 1	Automatic	Butanol, Cis-3-hexenol, Citronellol, Nonane, Phenethylamine, β-Pinene	Baseline VOC injection Recovery	6 s 15 s 110 s	6 Hz	71 (18)	15
Data set 2	Manual	Neopentane, Geranyl acetate, Allyl hexanoate, Butanoic acid, Propanoic acid, Valeric acid, Isovaleric acid	Baseline VOC injection Recovery	20 s 20 s 20 s	2 Hz	96 (25)	20

Tab. 3.1.: Description of the data sets used in this chapter. f_s corresponds to the number of frames per second (sampling frequency). P stands for the number of sensors and M for the number of sensing materials. N stands for the number of experiments per VOC. The recovery time (in the "time segmentation" column) corresponds to the recovery time which has been recorded. In the case of the manual acquisition (Date set 2), the actual recovery time is greater than 20 seconds to allow the complete desorption of each VOC.

To understand the problem, we propose to simplify Eq. (3.10) by considering the following model:

$$y_{pr}(t) = \begin{cases} 0 & \text{if } t < t_s \\ \theta_{pr}^{\text{eq}}\left(1 - e^{-\tau_{pr}^{a}(t-t_s)}\right) = y_{a,pr}(t) & \text{if } t \in [t_s, t_e] \\ y_{a,pr}(t_e)e^{-\tau_{pr}^{d}(t-t_e)} & \text{if } t > t_e \end{cases} \tag{3.12}$$

The model (3.12) no longer assumes any relation between the three new parameters $(\theta_{pr}^{\text{eq}}, \tau_{pr}^{a}, \tau_{pr}^{d})$. θ_{pr}^{eq} corresponds to the response at the equilibrium phase ($\theta_{pr}^{\text{eq}} = \frac{\nu_{pr} z_{pr}}{d_{pr}+z_{pr}}$). τ_{pr}^{a} stands for the constant related to adsorption ($\tau_{pr}^{a} = z_{pr} + d_{pr}$) and τ_{pr}^{d} for the constant related to desorption ($\tau_{pr}^{d} = d_{pr}$).

The fit of the model (3.12) performs much better. In Figure 3.6, we report the fits for 6 different VOCs taken from Data set 1 (see Table 3.1).

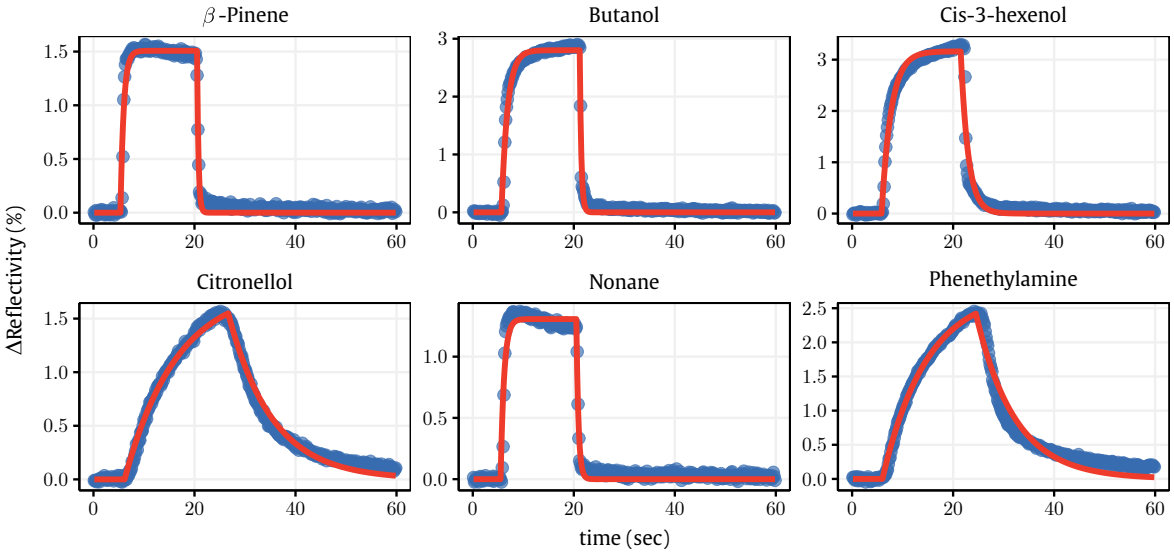


Fig. 3.6.: Fit of the model (3.12) for 6 VOCs from Data set 1. Blue points corresponds to the raw measurement of one chemical sensor. Red lines correspond to the fits. The time series have been truncated to 60 seconds for making easier the visualization (each experiment in Data set 1 lasts for $\sim$ 2 min).

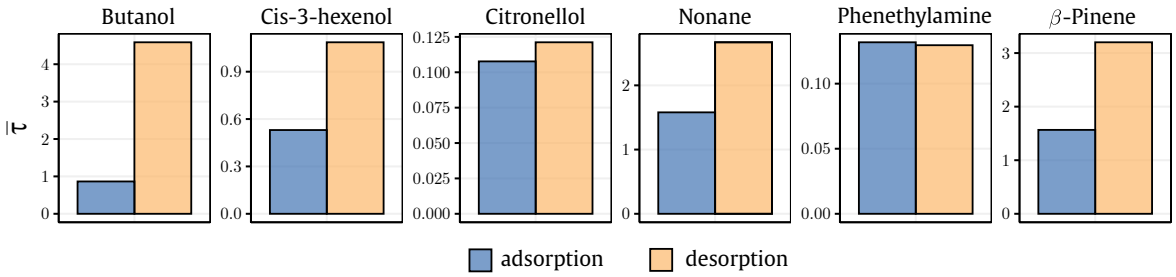


Fig. 3.7.: Estimated adsorption (blue, left bar) and desorption (yellow, right bar) constants from the model (3.12). For each VOC, the estimated parameters have been averaged over the P chemical sensors and the N measurements from Data set 1. According to the Langmuir model, the adsorption constant must be greater than the desorption constant. We observe the opposite in practice. The same observation holds for the data from Data set 2.

In Figure 3.7, we report a comparison between the two constants τ^a_{pr} and τ^d_{pr} estimated from real data. Figure 3.7 highlights a problem if one considers the model (3.10). According to the Langmuir model, the constant related to adsorption ($= z_{pr} + d_{pr}$) must be greater than the constant related to desorption ($= d_{pr}$). With the model (3.12), we show that it is the opposite in practice: τ^a_{pr} is almost always lower than τ^d_{pr}. It means that the kinetics of the binding reactions does not behave as expected by the Langmuir model. The same observation has been done on the data from Data set 2.

This empirical piece of evidence explains the difficulty to fit the model (3.10). In Figure 3.7, we observed that $\tau^a_{pr} \leqslant \tau^d_{pr}$. We can relate τ^a_{pr} and τ^d_{pr} to the Langmuir parameters, meaning that $\tau^a_{pr} = z_{pr} + d_{pr}$ and $\tau^d_{pr} = d_{pr}$. So, it comes that $z_{pr} + d_{pr} \leqslant d_{pr}$. Since all the parameters are non-negative, the best solution is then to put z_{pr} to 0. However, z_{pr} cannot be strictly null since z_{pr} has also a multiplicative effect on the global response (due to the term $\frac{\nu_{pr} z_{pr}}{d_{pr}+z_{pr}}$ in the model (3.10)). It is a balance which is complicated to manage for the optimization method explaining the presence of several local minima.

Limitations of the Langmuir model. The fit of the simplified model (Eq. (3.12)) performs well enough to predict the response of the chemical sensors (see Figure 3.6). However, the Langmuir model predicts poorly the constants of the exponential terms. In Section 3.1.1, we reviewed some assumptions which are required to apply the Langmuir model. In addition to these assumptions, VOC concentration over time must resemble a gate function to formulate this kinetic model. All of these assumptions can be violated in practice. Some of these violations can be highlighted only by carrying out additional measurements. Others can be directly stressed by the fits in Figure 3.6 by zooming in on some parts of the responses.

First, in Figure 3.8a, we highlight that the VOC concentration may not resemble a gate function, despite the use of an automatic valve in the case of Data set 1. When we measure VOCs from the headspace of a vial, this headspace is assumed to be stable meaning that the vaporization of the VOC is at the equilibrium phase. By measuring, we take some molecules of this vapor phase which results in a disruption of this equilibrium phase. If the vaporization is fast enough compared to the injection speed, this disruption can be rapidly compensated

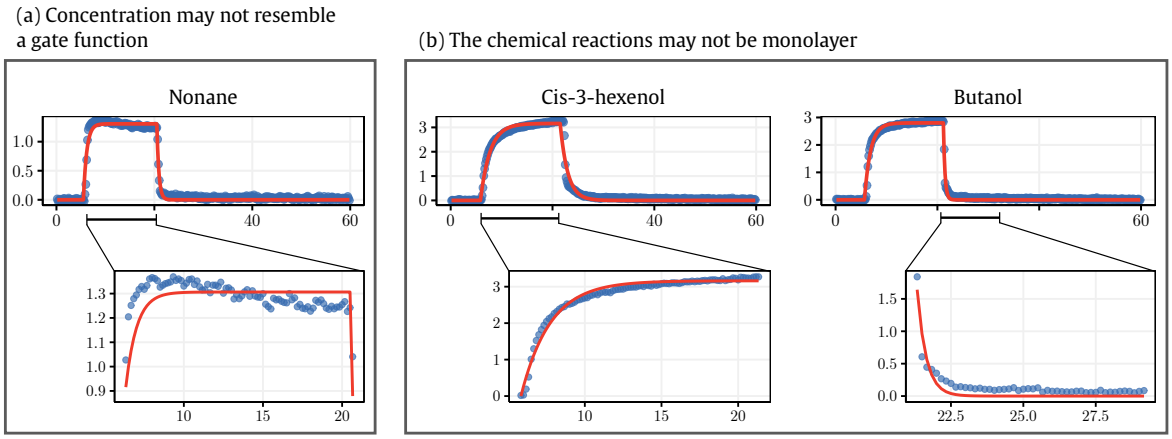


Fig. 3.8.: Data misfit in some parts of the responses may indicate the violations of some assumptions required for applying the Langmuir model. The data used are from Data set 1. (a) illustrates that concentration may not resemble a gate function and (b) illustrates that chemical reactions may not be monolayer.

in the vial. In this case, we can assume that the injected VOC concentration is constant. If the vaporization is not fast enough, the VOC concentration will not be constant over injection time and will decrease. Unfortunately, vaporization speed depends on the VOC. In the case of Nonane (see Figure 3.8a), the decrease of the plateau value indicates that the VOC concentration is decreasing during the VOC injection.

Second, in Figure 3.8b, we highlight that the response is not completely explained by a single exponential term, for both adsorption and desorption phases. It could be explained by another interaction than the one expected by the Langmuir model (see Eq. (3.3)). There are at least two possible reasons for that and they both rely on the existence of an additional chemical reaction. One reason is that the monolayer assumption may be too strong. The monolayer assumption implies that a site of a sensing material can be occupied only by a single molecule. In practice, it is possible that several molecules bind to a single site, leading to several layers. In this multilayer case, it could be interesting to derive an extension of the Langmuir model, starting off by considering two layers. However, there is currently another reason which is just as likely as the multilayer assumption. This other reason relies on an interaction between the gold layer and the VOC. In this case, the VOC binds not only to the sensing material but also to the gold surface. To reduce this possible interaction, Aryballe is currently working on a method called passivation. The idea is to protect the gold surface by covering it with a compound, thus limiting the possible interactions between the VOCs and gold.

In the current state of the technology, we do not increase the complexity of the model since the model (3.12) is still a good approximation of the response.

3.2.3 Feature extraction

Features. From the response of each chemical sensor, we generally extract one or two features per chemical sensor. The vector built with the P chemical sensors is called “signature” in the eNose community. There are a lot of possible features which can be extracted from the response and we explore some of them.

All these features could be combined to increase the information contained in the signature. However, the model (3.10) teaches us something about the response, even if this model has not been completely validated on real data. The response can likely be described with only three parameters $(\nu_{pr}, z_{pr}, d_{pr})$. Consequently, it may be redundant to use the entire time series or the combination of a lot of "different" features. In the following, we consider only one feature per chemical sensor.

First, we consider as possible features the parameters estimated from the model (3.12): θ_{pr}^{eq}, τ_{pr}^{a} and τ_{pr}^{d}. We saw that the model well describes real data but are the parameters enough discriminative to tell two VOCs apart?

Second, we consider four additional features which we can encounter in the eNose literature [Yan+15]. These features are extracted directly from the response and do not require any model fitting. Three of them are based on the integration of the response: the integration of the adsorption phase (from the beginning of the VOC injection to the beginning the plateau[5]), the integration of the desorption phase (from the beginning of the recovery to the end of the desorption[5]) and the integration of the whole response (from the beginning of the VOC injection to the end of the desorption). Finally, we also consider the steady state by averaging the plateau (from 3s before the beginning of the recovery to the beginning of the recovery). These four additional features are then based both on the transitory phases (adsorption and desorption) and the stationary phase (steady state). These features are redundant with the three parameters estimated from the model (3.12). For instance, θ_{pr}^{eq} and the steady state must be highly correlated since they represent the same quantity (the response at equilibrium). However, the parameters θ_{pr}^{eq}, τ_{pr}^{a} and τ_{pr}^{d} require fitting a model while the others are much easier to obtain. So, this comparison is interesting to justify or not the use of a model for extracting features.

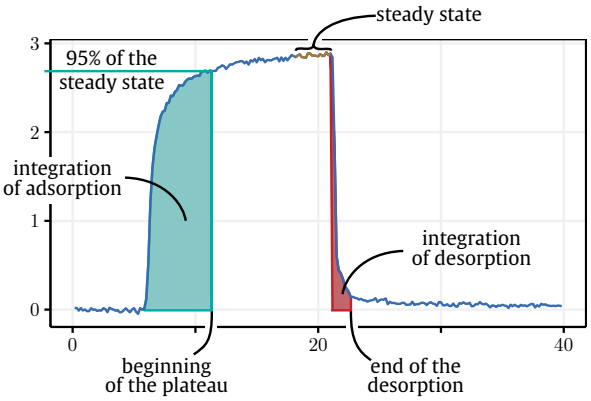


Fig. 3.9.: Features derived from the responses: the integration of adsorption, the integration of desorption and the steady state. The integration of the whole response (from the beginning of the VOC injection to the end of desorption) is not represented.

[5] We define as beginning of the plateau the time from which the response is greater than 95% of the steady state. We define as end of the desorption the time from which the response is lower than 5% of the steady state during recovery. These two times are computed on the average chemical sensor for a given measurement in order to have a fixed time range for all the chemical sensors. Another solution could be to define the integration times for each chemical sensor. However, in this case, the integration times may be different from one chemical sensor to another.

All these features may be influenced by the VOC concentration. To get rid of the concentration parameter, we normalize each signature by dividing by its ℓ_2-norm. It implicitly assumes that each feature scales linearly with VOC concentration[6].

Criterions. A fundamental aspect of a feature regarding classification is that it must be discriminative. To assess the discriminative power of a feature, we propose to use two criterions: the classification rate and the Fisher's ratio. The classification rate informs us about the ability of the feature to tell the VOCs apart. We estimate the classification rate by performing a 5-fold cross validation and by using a simple linear classifier, namely a linear Support-Vector Machine (SVM) [CV95]. Fisher's ratio [Fis36] informs us about the class separability of a given feature ([Bis06], p191). Class separability is an important characteristic of any feature. Ideally, the value of a feature is the same for samples belonging to the same class while it is as different as possible for samples belonging to different classes. In other words, distances between samples from the same class must be small while distances between samples from different classes must be large. Let us define the Fisher's ratio by noting $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ some data containing the samples of R VOCs, $\boldsymbol{\mu}_r \in \mathbb{R}^P$ the average sample of the VOC r and $\boldsymbol{\mu} \in \mathbb{R}^P$ the mean of $\boldsymbol{X}$. We note N_r the number of samples for the VOC r. Let us define two matrices ([Bis06], p191):

$$\boldsymbol{S}_w = \sum_{r=1}^{R} \sum_{n \in \text{VOC } r} (\boldsymbol{x}_n - \boldsymbol{\mu}_r)(\boldsymbol{x}_n - \boldsymbol{\mu}_r)^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{S}_b = \sum_{r=1}^{R} N_r (\boldsymbol{\mu}_r - \boldsymbol{\mu})(\boldsymbol{\mu}_r - \boldsymbol{\mu})^{\mathrm{T}}$$

$\boldsymbol{S}_w$ and $\boldsymbol{S}_b$ are referred to as the within-class covariance matrix and the between-class covariance matrix. $\boldsymbol{S}_w$ evaluates how far from each other are the samples belonging to a same class. The greater the within-class covariance is, the more spread-out the classes are. $\boldsymbol{S}_b$ evaluates how far from each other are the different classes. The greater the between-class covariance is, the farther from each other the different classes are. A good feature has a small within-class covariance and a high between-class covariance. The Fisher's ratio is computed by combining the two matrices and is equal to $\text{Tr}(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b)$[7]. The greater the Fisher's ratio is, the better.

Results. For each feature, we first reduce the dimensionality of the data. The reduction is based on PCA. Given the small number of samples used here, we reduce the dimensionality to 3, by projecting the data onto the first 3 Principal directions (see Figure 3.10). Both the classification rate and the Fisher's ratio are computed in this subspace. Results are reported in Table 3.2.

All the features perform equally well according to the classification rate. However, the Fisher's ratio shows some differences amongst the features and reveals that the features extracted

[6] Note $\tilde{\boldsymbol{x}} \in \mathbb{R}^P$ the unnormalized signature of a VOC and c the VOC concentration. If each feature in $\tilde{\boldsymbol{x}}$ scales linearly with c then $\tilde{\boldsymbol{x}} = c\boldsymbol{x}$ with $\boldsymbol{x}$ the concentration-free version of $\tilde{\boldsymbol{x}}$. Consequently, by taking the ℓ_2-norm of $\tilde{\boldsymbol{x}}$, we have $\|\tilde{\boldsymbol{x}}\|_2 = c\,\|\boldsymbol{x}\|_2$. Hence, $\frac{\tilde{\boldsymbol{x}}}{\|\tilde{\boldsymbol{x}}\|_2} = \frac{\cancel{c}\boldsymbol{x}}{\cancel{c}\|\boldsymbol{x}\|_2} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$

[7] $\text{Tr}(\cdot)$ is the Trace operator, meaning the sum of the diagonal terms of a matrix.

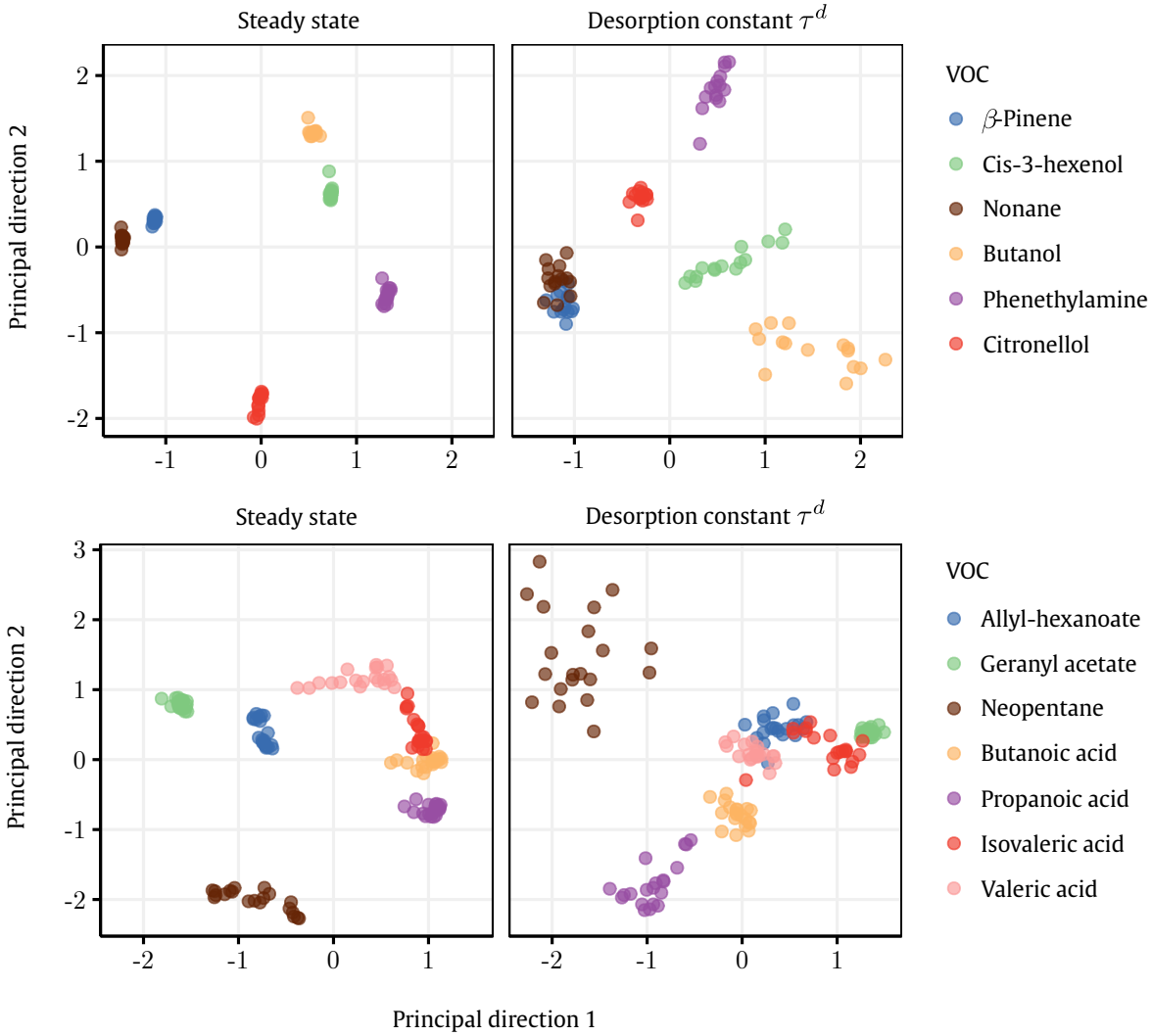


Fig. 3.10.: Projection of the data onto the two first Principal directions for Data set 1 (top) and Data set 2 (bottom). Left figures use the steady state as feature while right figures use τ^d as feature. Maps are centered and scaled for visual consideration.

directly from the response perform better. The parameters extracted from the model perform poorly according to class separability, especially the desorption constant τ^d. The best features according to the Fisher's ratio are the classical steady state and the integration of the whole response. Consequently, we did not find a clear advantage of using the features derived from a model.

We also tested the discriminative power of a combination of two different features. Since the classification was already almost perfect for all the features and for both data sets, the classification rate cannot tell us if there is a benefit of using several features instead of one. The Fisher's ratio shows little or no improvement by combining two features instead of using the best one (integration of the whole response or the steady state).

These results are obviously dependent on the data sets that we have used and especially on the VOCs studied. It may exists other data sets which would have shown different results.

		From the model			Directly from the response			
		τ^a	τ^d	θ^{eq}	Integration			Steady
					Ads.	Des.	All	state
Data set 1	Class. rate (%)	99	98	100	100	100	100	100
	Fisher's ratio	189	66	2564	2207	698	3971	3038
Data set 2	Class. rate (%)	99	99	99	99	99	99	100
	Fisher's ratio	63	41	155	196	84	268	297

Tab. 3.2.: Discriminative power of each feature, for each data set. For the Fisher's ratio, the greater, the better. In bold, we highlight the best features according to the two criterions.

However, given the results that we have obtained, we will not use the Langmuir model to extract features in the rest of this thesis. In the next chapters, we will focus on features extracted directly from the response such as the steady state or features based on integration.

As a final remark, it is interesting to note that the Fisher's ratios are much higher for Data set 1 than for Data set 2, whatever the feature. These differences could be explained by the acquisition method. Data set 1 has been acquired using an automatic valve (see Chapter 2) while Data set 2 has been acquired by hand. The latter can be affected by several variations caused by the human operator (*e.g.* the way the vial is brought to the instrument is never strictly identical), explaining a larger within-class covariance. However, the Fisher's ratio is not only dependent on the within-class covariance but also the between-class covariance. This covariance is greatly influenced by the choice of the VOCs and Data set 2 contains classes which may be less separable (especially the four VOCs from the acids family). To better quantify the influence of the automatic valve, we can look at the trace of the within-class covariance matrix for a given feature. This value corresponds to the total variance inside the classes. If we select the steady state as feature, we find that Data set 2 shows clusters which are $\sim$24 times more spread than the clusters from Data set 1. This observation is intuitive but can be important if one studies less separable classes than those we studied. However, the value 24 must be taken with caution since it depends on the feature and the two data sets are not only differing on the acquisition method but also on: the VOCs used, the eNose tested (not the same sensing materials), the environmental conditions, etc...

3.2.4 Conclusion

The model proposed in Section 3.1 allowed us to understand the different physical and chemical phenomena which are at the heart of the technology used by Aryballe's eNose. However, we showed with real data that this model presents weaknesses and that a simplification was useful bo better fit real responses.

We discussed about the possibility of using the parameters of the model as new features for classification. We showed that these parameters perform well according to classification performance. However, they perform less than more classical features according to a criterion based on class separability. We found that the classical steady state and the integration of the whole response perform significantly better.

In conclusion, the modeling of the response of Aryballe's eNose did not allow us to find better features for classification purpose. However, the Langmuir model may be useful for another task than feature extraction: Source Separation. For now, the formulation of the model is based on the assumption that we measure a single VOC. In the following, we modify the Langmuir model to take into account a gas mixture. This modification leads us to a non-linear mixture model for the equilibrium phase.

3.3 A non-linear mixture model

This section has been published in the Proceedings of the SAM conference in 2018[8]:

[hal-01802358] P. Maho, P. Comon, and S. Barthelmé
"*Non-linear source separation under the Langmuir model for chemical sensors*"
In: *10th IEEE Workshop on Sensor Array and Multichannel Signal Processing*, 2018, Sheffield (United Kingdom)

In Section 3.1, we described a model of the response of a chemical sensor over time. This model is the result of the combination of a chemical model (the Langmuir model) and a physical model describing the SPRi. However, the chemical model makes a strong hypothesis: the injected VOC is pure. In practical applications, it is likely that most of the gases detected by an eNose will be gas mixtures rather than pure VOCs.

For non-specific chemical sensors, we saw in Introduction (Chapter 1) that each VOC composing the mixture will interact with the chemical sensor. So the response of a non-specific chemical sensor to a gas mixture is then a non-trivial combination between VOC signatures and VOC concentrations. A solution to make the task easier would be to use specific chemical sensors which would respond only to a given VOC. However, the use of non-specific chemical sensors by eNoses is actually a requirement to detect a broad range of different VOCs (a specific chemical sensor would theoretically detect only a single VOC).

A fundamental question is then: how can we relate the response of Aryballe's eNose to the gas mixture (concentrations and signatures)? This question is fundamental for at least two reasons. First, it is simply important to understand how the measure of the instrument is generated not even in the pure VOC case (like in the previous section) but also in the more realistic case of a gas mixture. Beyond the sole understanding, it is also crucial for practical applications. In practice, one could use the global response of the eNose to a gas mixture as another signature that the instrument has to learn, just like the signature of a pure VOC. To illustrate this fact, we can use a well-known example of gas mixtures: perfumes. A perfume is sometimes composed of several hundreds of different VOCs, which is probably one of the most difficult mixture to study. A perfume can likely be identified by knowing which VOC is composing it and at which concentration. These differences in the compositions of two

[8] It is one of the first published works of this thesis, and it relies on a study made during the first PhD year. It details a non-linear model for describing the response of the instrument in presence of a gas mixture. In Chapter 6, we present another study about gas mixtures and we will see that a linear approximation performs actually well on real data. This study in Chapter 6 is subsequent to the work presented here.

different perfumes will likely generate two different signatures with an eNose. Therefore, a solution to identify these two perfumes could be to learn to differentiate them with Machine Learning algorithms, just like if they were two pure VOCs. However, this solution will likely fail if the composition of the perfume changes. For instance, it is well-known that a perfume changes over time due to the exhaustion of the most volatile compounds. In this case, the results will be greatly dependent on the data used for the learning phase. A more advantageous solution (but also much more complicated) would be to directly identify the composition of a perfume, meaning to decompose the measurement[9] according to VOC signatures and VOC concentrations. In the Signal Processing community, this is referred to as **Souce Separation** or **Unmixing**. To be able to decompose the measurement, we need then to formulate a model relating the response to concentrations and signatures.

In this section, we relax the hypothesis made in Section 3.1: we no longer measure a pure VOC but a gas mixture. We first define a theoretical non-linear mixture model describing how the equilibrium phase (*i.e.* the steady state) relates to concentrations and signatures. Then, we provide some theoretical properties of this new mixture model. Finally, we develop an algorithm to estimate both the signatures and the concentrations from measurements. The validation of this algorithm is based on numerical simulations.

3.3.1 Formulation

Langmuir mixture model. Let us assume that we measure a gas mixture composed of R VOCs, denoted by A_1, ..., A_R. To simplify the notation used to formulate the model, we first consider only one chemical sensor. To define the model, we need to reformulate the assumptions of the classical Langmuir model (see Section 3.1.1) [MB31]:

- the sensing surface is perfectly flat.
- the interaction is monolayer and each site can be occupied by only one VOC at a time.
- all the sites are equivalent.
- there is no interaction between the R VOCs.

Under these assumptions, each one of the R VOCs will interact with the sensing material S by a binding reaction. Considering all the reactions as independent of each other, there are R chemical reactions occurring on the surface:

$$A_r + S \underset{d_r}{\overset{a_r}{\rightleftharpoons}} A_r S \tag{3.13}$$

[9] It would be actually impossible to decompose a perfume according to VOC signatures and concentrations, at least with the current eNoses. For the decomposition to be possible, a basic assumption is that we have more sensing materials than there are VOCs actually mixing. Even if we consider simple perfumes with only one hundred VOCs, the actual number of sensing materials used by an eNose is far much smaller. For instance, in this thesis, Aryballe's eNose uses generally $\sim$20 different sensing materials (which is still much more than in other systems). Even if these sensing materials are replicated a large number of times to have more chemical sensors, it does not solve the problem. In fact, the information supported by two replicas of a same sensing material is likely much lower than the information supported by two different sensing materials. So, the research of new sensing materials is fundamental for these applications to be possible. This research is currently in progress [Hur+19; Gag+20]

where a_r and d_r respectively stand for the adsorption and desorption constants of VOC r.

Similarly to Section 3.1.1, we can use the notion of free/occupied sites to describe the chemical interactions. In the case of a gas mixture, each free site can be occupied by one (and only one) of the R VOCs. So, the number of occupied sites is no longer dependent on a single VOC as in the classical Langmuir model. In the model formulation, we need to consider that the number of occupied sites is now dependent on the R different VOCs. We can then rewrite Eq. (3.2) for describing the number of sites becoming occupied by the VOC r, taking into account the R VOCs:

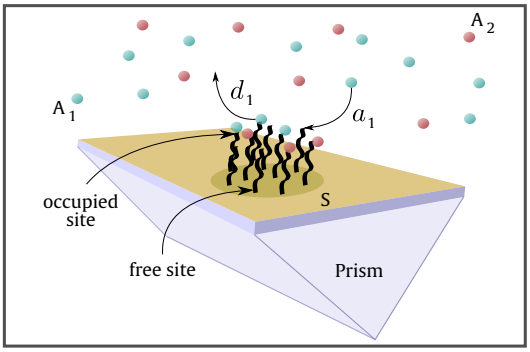


Fig. 3.11.: Free/occupied sites for a gas mixture.

$$\frac{\mathrm{d}[\mathrm{A}_r\mathrm{S}]_s(t)}{\mathrm{d}t} = a_r[\mathrm{A}_r](t)\big([\mathrm{S}]_s - \sum_{i=1}^{R}[\mathrm{A}_i\mathrm{S}]_s(t)\big) - d_r[\mathrm{A}_r\mathrm{S}]_s(t) \tag{3.14}$$

Similarly to Section 3.1.1, we can assume that $[\mathrm{A}_r](t)$ is constant during injection, of value c_r, and null outside injection. We can also easily derive a differential equation with respect to the fraction of occupied sites by the VOC r, $\theta_r(t) = \frac{[\mathrm{A}_r\mathrm{S}]_s(t)}{[\mathrm{S}]_s}$:

$$\frac{\mathrm{d}\theta_r(t)}{\mathrm{d}t} = a_r c_r\big(1 - \sum_{i=1}^{R}\theta_i(t)\big) - d_r\theta_r(t) \tag{3.15}$$

If we wait for a sufficiently long time, the system of chemical reactions will reach an equilibrium phase, meaning that $\frac{\mathrm{d}\theta_r(t)}{\mathrm{d}t} = 0$. In this case, Eq. (3.15) becomes:

$$a_r c_r\big(1 - \sum_{i=1}^{R}\theta_i^{\mathrm{eq}}\big) = d_r\theta_r^{\mathrm{eq}} \tag{3.16}$$

By summing over the VOCs, we can first isolate $\sum_{r=1}^{R}\theta_r^{\mathrm{eq}}$. We plug $k_r = \frac{a_r}{d_r}$:

$$\begin{aligned}
(3.16) \iff & \quad k_r c_r\big(1 - \sum_{i=1}^{R}\theta_i^{\mathrm{eq}}\big) &&= && \theta_r^{\mathrm{eq}} \\
\implies & \quad \sum_{r=1}^{R} k_r c_r\big(1 - \sum_{i=1}^{R}\theta_i^{\mathrm{eq}}\big) &&= && \sum_{r=1}^{R}\theta_r^{\mathrm{eq}} \\
\implies & \quad \sum_{r=1}^{R}\theta_r^{\mathrm{eq}} &&= && \frac{\sum_{r=1}^{R} k_r c_r}{1 + \sum_{r=1}^{R} k_r c_r}
\end{aligned} \tag{3.17}$$

By replacing Eq. (3.17) in Eq. (3.16):

$$\theta_r^{\text{eq}} = \frac{k_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} \tag{3.18}$$

The model (3.18) is an extension of the Langmuir model in the case of a gas mixture of R VOCs, derived by E. C. Markham and A. F. Benton [MB31].

Langmuir-SPRi mixture model. Eq. (3.18) corresponds to the chemical model but the response of Aryballe's eNose results also from other phenomena, namely SPRi (the transduction method). In Section 3.1.2, we described SPRi measurement as being proportional to the fraction of occupied sites. We recall that this approximation works if the thickness e_b of the medium formed by the adsorption of the VOC on the sensing material is small enough. We still assume that this assumption is valid in the case of a gas mixture. Therefore, we can write that:

$$\Delta R_r^{\text{eq}} = \frac{S}{L_z}[\text{S}]_s \frac{\partial n_r}{\partial c_r} m_r \times \theta_r^{\text{eq}} \tag{3.19}$$

with ΔR_r^{eq} the reflectivity (after baseline subtraction) at the equilibrium phase due to the adsorption of the VOC r. S and L_z are constants related to SPRi settings and $[\text{S}]_s$ is the initial number of free sites. We take the same notation as in the previous section by noting: $\alpha_{\text{spr}} = \frac{S}{L_z}$, $\phi = [\text{S}]_s$ and $\eta_r = \frac{\partial n_r}{\partial c_r} m_r$. We note $\tilde{x}_r$ the SPRi response during equilibrium phase for the VOC r:

$$\tilde{x}_r = \alpha_{\text{spr}} \phi \eta_r \times \frac{k_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} \tag{3.20}$$

However, we do not directly measure the fraction of occupied sites of each VOC r. Our measure is more global and is related to the total fraction of occupied sites. So the model (3.20) is not the final one.

To find a model leading to the global response $\tilde{x}$ that we actually measure, we make one more assumption. We already assumed that the thickness e_b of the medium is small enough to assume the linearity of the SPRi response. We assume that this assumption also implies that the medium will be seen as being "homogeneous" by the SPRi, so the individual responses, $\tilde{x}_r$, simply add up:

$$\tilde{x} = \alpha_{\text{spr}} \phi \times \frac{\sum_{r=1}^{R} k_r \eta_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} \tag{3.21}$$

We can notice that $\tilde{x}$ corresponds to the steady state feature that we used in Section 3.2.3 for classification purpose.

Finally, we propose to simplify the model (3.21) by noticing that a simple normalization can theoretically remove the contributions of α_{spr} and ϕ.

For that, let us assume that before the experiments with gas mixtures, we carry out a preliminary measurement with a pure VOC, called the VOC 0, at concentration c_0 having the parameters a_0 and d_0. For any VOC, one can reach a maximum concentration from which the response will no longer change (even if we increase over and over again the concentration) [Hom06] (p78). This is due to the fact that all the sites of the sensing material S are occupied at this boundary concentration. In other words, the fraction of occupied sites θ_0 equals 1 and is then no longer dependent on a_0, d_0 or c_0. We can show it mathematically. Let us assume that we are at the equilibrium phase of this preliminary measurement (so we have waited a sufficiently long time), and note θ_0^{eq} the fraction of occupied sites:

$$c_0 \gg \frac{d_0}{a_0} \Longrightarrow \theta_0^{\text{eq}} = \frac{a_0 c_0}{d_0 + a_0 c_0} = \frac{1}{\frac{d_0}{a_0 c_0} + 1} \approx 1$$

If $\theta_0^{\text{eq}} \approx 1$, then the SPRi response at equilibrium becomes simply $\tilde{x}_0 \approx \alpha_{\text{spr}} \phi \eta_0$, with $\eta_0 = \frac{\partial n_0}{\partial c_0} m_0$. So, theoretically, we can easily remove the influence of the parameters α_{spr} and ϕ on the response of any future measurement. To illustrate that, let us assume that we now measure the steady state value $\tilde{x}$ of a gas mixture. We can normalize $\tilde{x}$ using $\tilde{x}_0$:

$$x = \frac{\tilde{x}}{\tilde{x}_0} = \frac{\alpha_{\text{spr}} \phi}{\alpha_{\text{spr}} \phi \eta_0} \times \frac{\sum_{r=1}^{R} k_r \eta_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} = \frac{1}{\eta_0} \times \frac{\sum_{r=1}^{R} k_r \eta_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} \tag{3.22}$$

In practice, this trick would only require to perform several experiments with a chosen VOC (the VOC 0) and to identify this max concentration. The simple normalization provided by Eq. (3.22) is still theoretical and we did not try it in practice, but we assume in the following that we can apply it for removing the constants α_{spr} and ϕ (it makes the formulation easier). As a remark, η_0 is a known parameter since we know the VOC 0, so we can multiply the normalized response x by η_0. Therefore, the final model that we consider is the model after normalization and multiplication by η_0:

$$x = \frac{\sum_{r=1}^{R} k_r \eta_r c_r}{1 + \sum_{r=1}^{R} k_r c_r} \tag{3.23}$$

The model (3.23) corresponds to the theoretical response at the equilibrium phase of a gas mixture of R VOCs, but considering only one chemical sensor and one experiment. In practice, an eNose has an array of P chemical sensors. We make the assumption here that each one of the P chemical sensors has a different sensing material (or that we have averaged the replicas), so we are in the case $P = M$ (with M the notation for the number of sensing

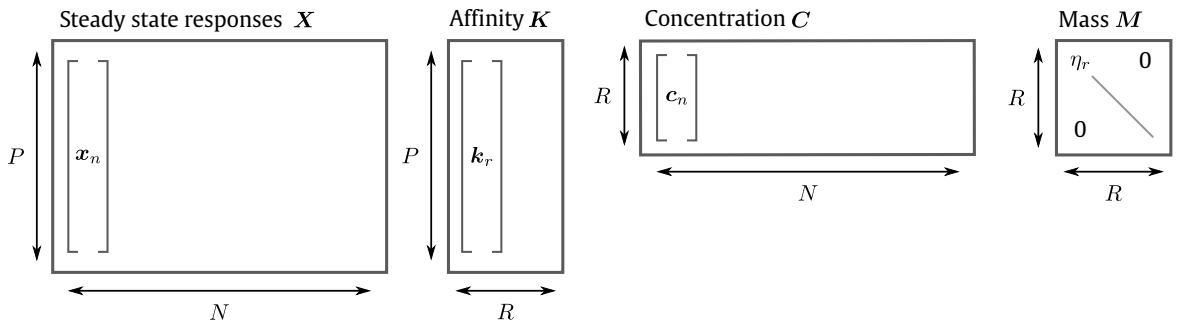


Fig. 3.12.: Graphical representation of the four matrices used in the model (3.25).

materials). In fact, the replicas of a sensing material are theoretically uninformative according to the Langmuir model (they share the same parameters). In practice, two replicas can however respond differently, but we currently do not know why (and especially how to model these variations). Several experiments are also carried out, say N experiments. To highlight these dimensions in the model (3.23), we note p and n the subscripts referring respectively to the chemical sensor and to the experiment index. The affinity constant k_r is dependent on the chemical sensor. The concentration c_r is dependent on the experiment (the proportion of each VOC may vary from one experiment to another). The model (3.23) becomes:

$$\boxed{x_{pn} = \frac{\sum_{r=1}^{R} k_{pr}\eta_r c_{rn}}{1 + \sum_{r=1}^{R} k_{pr} c_{rn}}} \tag{3.24}$$

with x_{pn} the steady state response of the chemical sensor p during experiment n. By considering all the experiments and all the chemical sensors, we can propose a matrix formulation of the model (3.24). Let us note:

- $\boldsymbol{x}_n \in \mathbb{R}^P$, the column vector containing the P steady state values of the P chemical sensors during the experiment n.
- $\boldsymbol{c}_n \in \mathbb{R}^R$, the column vector containing the R VOC concentrations during the mixture injection in the experiment n.
- $\boldsymbol{k}_r \in \mathbb{R}^P$, the column vector containing the P affinity constants of the P chemical sensors for the VOC r.
- $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N] \in \mathbb{R}^{P\times N}$, the measurement matrix containing the steady states.
- $\boldsymbol{C} = [\boldsymbol{c}_1, ..., \boldsymbol{c}_N] \in \mathbb{R}^{R\times N}$, the concentration matrix.
- $\boldsymbol{K} = [\boldsymbol{k}_1, ..., \boldsymbol{k}_R] \in \mathbb{R}^{P\times R}$, the affinity matrix.
- $\boldsymbol{M} \in \mathbb{R}^{R\times R}$, a diagonal matrix containing the R parameters η_r (related for instance to the mass).
- $\oslash$ the Hadamard (*i.e.* entrywise) division.

Then, the model (3.24) can be written as:

$$\boxed{\boldsymbol{X} = \boldsymbol{K}\boldsymbol{M}\boldsymbol{C} \boxslash (\mathbf{1}_P \mathbf{1}_N^{\mathrm{T}} + \boldsymbol{K}\boldsymbol{C})} \tag{3.25}$$

In Figure 3.12, we provide a graphical representation of the four matrices $\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{C}$ and $\boldsymbol{M}$. We referred to the model (3.25) as the **Langmuir-SPRi mixture model**. We did not find the description of this model in the literature so we provide in the next section some theoretical properties.

3.3.2 Theoretical results

In practice, $\boldsymbol{X} \in \mathbb{R}^{P \times N}$ is a known data matrix extracted from the steady states of N experiments using the Aryballe's eNose with P chemical sensors. In these N experiments, a gas mixture composed of R VOCs at different concentrations has been measured. The R VOCs have certain affinities with the P chemical sensors and this information is contained in $\boldsymbol{K} \in \mathbb{R}^{P \times R}$. The responses $\boldsymbol{X}$ are dependent on the concentration of the R VOCs during the N experiments and this information is contained in $\boldsymbol{C} \in \mathbb{R}^{R \times N}$. In addition, the responses are also dependent on some parameters only related to the R VOCs (*e.g.* mass) and this information is contained in the diagonal matrix $\boldsymbol{M} \in \mathbb{R}^{R \times R}$. We assume that the three matrices $\boldsymbol{K}, \boldsymbol{C}$ and $\boldsymbol{M}$ gave us the observed data matrix $\boldsymbol{X}$ according to the non-linear model (3.25). These three matrices carry important information about the VOCs and the mixtures, and $\boldsymbol{K}$ and $\boldsymbol{C}$ are particularly important for obvious reasons ($\boldsymbol{K}$ gives us the identity of each VOC while $\boldsymbol{C}$ quantifies the presence of each VOC in the mixtures). $\boldsymbol{K}, \boldsymbol{C}$ and $\boldsymbol{M}$ can be assumed to be known but, in practice, at least one of these matrices will be unknown and will need to be estimated from the data. We can distinguish three possible types of estimation problem:

- **Supervised unmixing** or **Regression**. $\boldsymbol{M}$ is known. Either $\boldsymbol{K}$ or $\boldsymbol{C}$ is unknown. Regression will be studied on real data in Chapter 6 using a simpler model (but which can be derived from the Langmuir-SPRi mixture model) and assuming that $\boldsymbol{K}$ is known.
- **Blind Source Separation**. $\boldsymbol{M}$ is known. Both $\boldsymbol{K}$ and $\boldsymbol{C}$ are unknown. In the next section, we develop an algorithm for estimating both $\boldsymbol{K}$ and $\boldsymbol{C}$ and we validate this algorithm on simulated data.
- **Fully Blind Source Separation**. $\boldsymbol{K}$, $\boldsymbol{C}$ and $\boldsymbol{M}$ are unknown.

These three problems are not on the same level of difficulty for a simple reason: the greater the number of parameters to estimate, the harder the problem. So Fully Blind Source Separation is clearly not trivial and may be impossible to solve (in particular, it may exist many different possible parameters $(\boldsymbol{K}, \boldsymbol{C}, \boldsymbol{M})$ giving the same data $\boldsymbol{X}$). In fact, we did not study this case and our results are limited to the Regression case and to the Blind case. So, in the following, we always assume that $\boldsymbol{M}$ is known. We would like to stress that this assumption is strong for a simple reason: $\boldsymbol{M}$ depends on the R VOCs so, assuming its knowledge implies that we know the VOCs which are mixing or at least that we know each parameter η_r which can be

limiting (we recall that $\eta_r = \frac{\partial n_r}{\partial c_r} m_r$, with m_r the VOC mass and $\frac{\partial n_r}{\partial c_r}$ the change in refractive index of the VOC with change in concentration).

Supervised unmixing

Either $\boldsymbol{K}$ or $\boldsymbol{C}$ is unknown and the rest of the parameters is assumed to be known.

If $\boldsymbol{K}$ is unknown, then we assume that we have more experiments than there are VOCs in the mixtures, *i.e.*, $N \geqslant R$. If $\boldsymbol{C}$ is unkwown, then we assume that we have more chemical sensors than there are VOCs in the mixtures, *i.e.*, $P \geqslant R$. These two conditions simply state that we have enough equations ($= P \times N$) compared to the number of unknown parameters ($= P \times R$ if $\boldsymbol{K}$ is unknown or $R \times N$ if $\boldsymbol{C}$ is unknown), which is a prerequisite to be able to estimate either $\boldsymbol{K}$ or $\boldsymbol{C}$.

To find identifiability conditions for $\boldsymbol{K}$ or $\boldsymbol{C}$, we rewrite the model (3.25) by noticing that it is separable across experiments:

$$\begin{aligned} (3.25) &\iff \forall n, \quad \boldsymbol{x}_n \boxdot (\mathbf{1}_P + \boldsymbol{K}\boldsymbol{c}_n) = \boldsymbol{K}\boldsymbol{M}\boldsymbol{c}_n \\ &\iff \forall n, \quad \boldsymbol{x}_n = (\boldsymbol{K}\boldsymbol{M} - \text{diag}(\boldsymbol{x}_n)\boldsymbol{K})\boldsymbol{c}_n \end{aligned} \tag{3.26}$$

with $\boxdot$ the Hadamard (*i.e.* entrywise) product.

It is also separable across chemical sensors, noting $\boldsymbol{x}_p^{\mathrm{T}} \in \mathbb{R}^N$ the p^{th} row of $\boldsymbol{X}$ and $\boldsymbol{k}_p^{\mathrm{T}} \in \mathbb{R}^R$ the p^{th} row of $\boldsymbol{K}$:

$$(3.25) \iff \forall p, \quad \boldsymbol{x}_p^{\mathrm{T}} = \boldsymbol{k}_p^{\mathrm{T}}(\boldsymbol{C}\text{diag}(\boldsymbol{x}_p^{\mathrm{T}}) - \boldsymbol{M}\boldsymbol{C}) \tag{3.27}$$

Lemma 1. *Let $\boldsymbol{H}_n \overset{def}{=} \boldsymbol{K}\boldsymbol{M} - diag(\boldsymbol{x}_n)\boldsymbol{K}$. Then $\boldsymbol{c}_n$ is identifiable if and only if $\boldsymbol{H}_n$ is full rank. Similarly, let $\boldsymbol{G}_p \overset{def}{=} \boldsymbol{C}diag(\boldsymbol{x}_p^T) - \boldsymbol{M}\boldsymbol{C}$. Then $\boldsymbol{k}_p^T$ is identifiable if and only if $\boldsymbol{G}_p$ is full rank.*

Proof. The proof is straightforward: assume there are two solutions $\boldsymbol{c}_n$ and $\boldsymbol{c}'_n$. Then $\boldsymbol{H}_n\boldsymbol{c}_n = \boldsymbol{x}_n$ and $\boldsymbol{H}_n\boldsymbol{c}'_n = \boldsymbol{x}_n$. By subtraction, $\boldsymbol{H}_n(\boldsymbol{c}_n - \boldsymbol{c}'_n) = \mathbf{0}_P$. Since $\boldsymbol{H}_n$ is full rank, we have necessarily $(\boldsymbol{c}_n - \boldsymbol{c}'_n) = \mathbf{0}_P$. The proof for $\boldsymbol{k}_p^{\mathrm{T}}$ is the same. □

Blind Source Separation

$\boldsymbol{K}$ and $\boldsymbol{C}$ are unknown and $\boldsymbol{M}$ is assumed to be known.

In this case, it is easy to see that we have $P \times N$ equations compared to $P \times R + N \times R$ unknown parameters. A prerequisite without any additional information (such as constraints) is that $\frac{P \times R + N \times R}{P \times N} \leqslant 1$, which can be rewritten as $\frac{R}{N} + \frac{R}{P} \leqslant 1$. Consequently, it is interesting to note that the condition $P \geqslant R$ and $N \geqslant R$ is not enough to satisfy this prerequisite (*e.g.* take $P = N = R$). In the following, we assume that $P \geqslant 2R$ and $N \geqslant 2R$, ensuring that we

have enough equations compared to the number of unknown parameters[10] (however other situations work too such as $P \geqslant R$ and $N \geqslant 3R$, etc...).

Some limiting cases. The Langmuir-SPRi mixture model in Eq (3.25) is a non-linear model. For the three estimation problems introduced before, a non-linear model is by nature more complicated to handle than a linear model. However, it is interesting to note that the Langmuir-SPRi mixture model can be simplified considering some specific experimental scenarios. In these scenarios, the problems may be easier or harder.

- Low concentration or low affinity regime. When we deal with an application where low concentrations are expected, this leads to a simpler version of the model (3.25). Let us assume $\boldsymbol{C} \ll 1$ (*i.e.* each coefficient in $\boldsymbol{C}$ is much smaller than 1) and $\boldsymbol{K} \approx O(1)$ (*i.e.* each coefficient in $\boldsymbol{K}$ is dominated by 1), we can simplify (3.25) by a 0-order Taylor expansion:

$$\boldsymbol{X} \approx \boldsymbol{KMC} \approx \boldsymbol{K\tilde{C}} \tag{3.28}$$

  with $\tilde{\boldsymbol{C}} = \boldsymbol{MC}$ (recall that $\boldsymbol{M}$ is diagonal).

  The model (3.28) is the well-known linear mixture model. For instance, identifiability conditions in the Blind case have been addressed in the literature (see for example some conditions for Non-negative Matrix Factorization [HSS14] or the conditions of Independent Component Analysis [Com94]). Since the roles of $\boldsymbol{K}$ and $\boldsymbol{C}$ are symmetrical in the model (3.25), the approximation also works for: $\boldsymbol{K} \ll 1$ and $\boldsymbol{C} \approx O(1)$. In Chapter 6, we use the model (3.28) on real data obtained from robotic applications. Finally, we can note that the linear approximation works even better in the case of low affinity and low concentration regime.

- Saturation regime. At the opposite end, if we work with really high concentrations or with VOCs having a really high affinity with the sensor array, then $\boldsymbol{KC} \gg 1$ (*i.e.* each coefficient in $\boldsymbol{KC}$ is much greater than 1) and (3.25) becomes harder:

$$\boldsymbol{X} \approx \boldsymbol{KMC} \oslash \boldsymbol{KC} \tag{3.29}$$

  Why is the model (3.29) harder than the model (3.25)? It is a matter of identifiability. The model (3.29) is clearly suffering from an indetermination.

  Let $(\boldsymbol{D}_1, \boldsymbol{D}_2) \in \mathbb{R}^{P\times P} \times \mathbb{R}^{N\times N}$ be two diagonal matrices such that none of the diagonal terms is null, then:

$$\boldsymbol{D}_1\boldsymbol{KMCD}_2 \oslash \boldsymbol{D}_1\boldsymbol{KCD}_2 = \boldsymbol{KMC} \oslash \boldsymbol{KC} \tag{3.30}$$

[10] In the paper [MBC18] that we published in the SAM conference (2018), we have made a mistake by considering $P \geqslant R$ and $N \geqslant R$, at least for the theoretical part. For the numerical simulations, we had chosen $P = N = 100$ while R was max 15. So, for the numerical results, the number of equations was actually greater than the number of unknown parameters.

The indetermination (3.30) means that, for instance, each concentration vector $\boldsymbol{c}_n \in \mathbb{R}^R$ for an experiment n can be multiplied by a factor $\epsilon \in \mathbb{R}^*$ while giving the same result. Fortunately, this version of (3.25) is unlikely to appear in practice since it would be useless to put the sensor in a saturation regime.

- VOCs with the same parameter η_r. As a final limiting case, let us consider some VOCs having the same parameter $\eta_r = \eta$. In this case, $\boldsymbol{M}$ is simply $\eta\boldsymbol{I}$, so (3.25) becomes:

$$\frac{1}{\eta}\boldsymbol{X} = \boldsymbol{KC} \oslash (\mathbf{1}_P\mathbf{1}_N^{\mathrm{T}} + \boldsymbol{KC}) \tag{3.31}$$

  This particular model falls into a class known as Post Non-Linear mixtures [TJ99; AJ05]. This kind of models means that the non-linearity can be seen as a non-linear function that we have applied to the linear model. Therefore, by simply applying the inverse function to the data, we can then assume the linear model. In our case, the component-wise inversion $g(x) = \frac{x}{1-x}$ would allow us to recover the linear model $\boldsymbol{X}_g = \boldsymbol{KC}$. The model (3.31) is interesting but rather unrealistic in an application prospect, as it considerably limits the set of VOCs that can be mixed together.

  Finally, it is interesting to note that the model (3.31) corresponds to the Langmuir mixture model if we do not consider the SPRi measurement (the matrix $\boldsymbol{M}$ comes only from the SPRi model). In other words, the Langmuir mixture model can be seen as a hidden linear model. Therefore, it may be easier to handle than the Langmuir-SPRi mixture model that we consider.

Identifiability. First, a usual trivial indetermination in a Blind Source Separation framework is the source permutation, at least in the simple case where the model is linear. To illustrate this indetermination, let us assume the linear model $\boldsymbol{X} = \boldsymbol{KC}$. Then, for any invertible matrix $\boldsymbol{Q} \in \mathbb{R}^{R\times R}$, we have $\boldsymbol{X} = \boldsymbol{KQ}^{-1}\boldsymbol{QC} = \boldsymbol{KC}$ so the solutions $(\boldsymbol{KQ}^{-1}, \boldsymbol{QC})$ and $(\boldsymbol{K}, \boldsymbol{C})$ are equivalent. By constraining the problem, we can limit the set of invertible matrices $\boldsymbol{Q}$ causing the indetermination to the set of permutation and scaling matrices (see for instance the indeterminations in the case of Independent Component Analysis [Com94]). Permutation means that we can switch, for instance, the first and second columns of $\boldsymbol{K}$ while switching the first and second rows of $\boldsymbol{C}$ and we will still have the same resulting matrix $\boldsymbol{X}$.

It is clear that the model (3.25) suffers from a scaling indetermination, meaning that for any diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{R\times R}$, the solutions $(\boldsymbol{KD}^{-1}, \boldsymbol{DC})$ and $(\boldsymbol{K}, \boldsymbol{C})$ are equivalent. However, it is interesting to note that the model (3.25) does not suffer from the permutation indetermination, subject to a weak assumption.

Indeed, let $\boldsymbol{Q} \in \mathbb{R}^{R\times R}$ be an invertible matrix, and keeping in mind that $\boldsymbol{M}$ is diagonal, we have:

$$\boldsymbol{K}\boxed{\boldsymbol{QMQ}^{-1}}\boldsymbol{C} \oslash (\mathbf{1}_P\mathbf{1}_N^{\mathrm{T}} + \boldsymbol{KQQ}^{-1}\boldsymbol{C}) \quad = \quad \boldsymbol{KMC} \oslash (\mathbf{1}_P\mathbf{1}_N^{\mathrm{T}} + \boldsymbol{KC})$$

If and only if,

$$\begin{array}{lll} & \boldsymbol{QMQ}^{-1} = \boldsymbol{I}_R \\ \Leftrightarrow & \boldsymbol{QM} = \boldsymbol{MQ} \\ \Leftrightarrow \quad \forall i,j & (\boldsymbol{QM})_{ij} = (\boldsymbol{MQ})_{ij} \\ \Leftrightarrow \quad \forall i,j & \sum_{r=1}^{R} q_{ir}m_{rj} = \sum_{r=1}^{R} m_{ir}q_{rj} \\ \Leftrightarrow \quad \forall i,j & q_{ij}m_{jj} = m_{ii}q_{ij} \\ \Leftrightarrow \quad \forall i \neq j & q_{ij}m_{jj} = m_{ii}q_{ij} \\ \Leftrightarrow \quad \forall i \neq j & q_{ij} = 0 \quad \text{or} \quad m_{jj} = m_{ii} \end{array} \tag{3.32}$$

Thus, if we assume that each VOC r of the mixture has a different parameter η_r, meaning that $\forall i \neq j,\ m_{jj} \neq m_{ii}$, then the only trivial indetermination which remains is the scaling ambiguity.

In conclusion, the identifiability of the model (3.25) can be only obtained up to a diagonal matrix (scaling indetermination). However, proving the identifiability of the model (3.25) is difficult and we only provide a necessary condition.

Proposition 1 (Necessary condition)**.** *If $\boldsymbol{c}_n$ (resp. $\boldsymbol{k}_p^T$) is identifiable $\forall n$ (resp. $\forall p$), then $\boldsymbol{K}$ (resp. $\boldsymbol{C}$) is full rank, or $\boldsymbol{M}$ is not proportional to the Identity matrix.*

Proof. Let us prove the statement for $\boldsymbol{K}$, since a dual reasoning holds for $\boldsymbol{C}$. The proof is by contradiction. Assume $\text{rank}(\boldsymbol{K}) \neq R$ and $\boldsymbol{M} \propto \boldsymbol{Id}$, then $\exists \boldsymbol{v} \in \mathbb{R}^R \backslash \{\boldsymbol{0}_R\}$ such that $\boldsymbol{Kv} = \boldsymbol{0}_P$. Hence $\text{diag}(\boldsymbol{x}_n)\boldsymbol{Kv} = \boldsymbol{0}_P$. On the other hand, because $\boldsymbol{M} \propto \boldsymbol{Id}$, $\boldsymbol{KMv} = \boldsymbol{0}_P$. Hence $\boldsymbol{H}_n\boldsymbol{v} = \boldsymbol{0}_P$, which prevents identifiability from Lemma 1. □

3.3.3 An algorithm in the Blind Source Separation framework

In the previous section, we did not find a sufficient condition for guaranteeing the identifiability of the Langmuir-SPRi mixture model, when both $\boldsymbol{K}$ and $\boldsymbol{C}$ are unknown (recall that $\boldsymbol{M}$ is known). However, in this section, we study to what extent the model inversion (*i.e.* estimation of $\boldsymbol{K}$ and $\boldsymbol{C}$ from $\boldsymbol{X}$) is possible in simulation.

In the following, we note the Langmuir-SPRi mixture model as:

$$\mathscr{L}(\boldsymbol{K},\boldsymbol{C}) = \boldsymbol{KMC} \oslash (\boldsymbol{1}_P\boldsymbol{1}_N^{\mathrm{T}} + \boldsymbol{KC})$$

Optimization problem. We need to first define a cost function, measuring the error between some data $\boldsymbol{X} \in \mathbb{R}^{P\times N}$ and the data predicted by the model $\mathscr{L}(\boldsymbol{K},\boldsymbol{C})$ using the parameters $\boldsymbol{K} \in \mathbb{R}^{P\times R}$ and $\boldsymbol{C} \in \mathbb{R}^{R\times N}$. A classical way to do that is to consider the least squares cost function:

$$\Upsilon(\boldsymbol{K},\boldsymbol{C}) = \|\boldsymbol{X} - \mathscr{L}(\boldsymbol{K},\boldsymbol{C})\|_F^2 \tag{3.33}$$

where $\|\cdot\|_F$ is the Frobenius norm (square root of the sum of the squared terms of a matrix).

By noticing that $\boldsymbol{K}$ and $\boldsymbol{C}$ are non-negative (recall that $\boldsymbol{K}$ corresponds to chemical affinities and $\boldsymbol{C}$ to VOC concentrations, so all these terms are "naturally" non-negative), we need to solve the following optimization problem:

$$\hat{\boldsymbol{K}}, \hat{\boldsymbol{C}} = \underset{\boldsymbol{K}\geqslant 0, \boldsymbol{C}\geqslant 0}{\arg\min}\ \Upsilon(\boldsymbol{K}, \boldsymbol{C}) \tag{3.34}$$

where $\boldsymbol{K} \geqslant 0$ means that all the terms of $\boldsymbol{K}$ are non-negative.

The previous study about some limiting cases of the model (3.25) shows that if the terms of the product $\boldsymbol{KC}$ becomes too large (saturation regime), then the model suffers from an indetermination. This indetermination may lead to unstability concerning the estimation of the parameters. To avoid this "non-identifiable" area, we propose to constrain the product $\boldsymbol{KC}$ assuming that we are far from the saturation regime (*i.e.* high affinity or high concentration). A way to do that is to constrain the max-norm: $\|\boldsymbol{KC}\|_{\max} \stackrel{\text{def}}{=} \max_{ij}(|(\boldsymbol{KC})_{ij}|)$, meaning the maximum value of the product $\boldsymbol{KC}$. We assume that the max-norm is lower than Ω (a parameter to define), leading to the new optimization problem:

$$\hat{\boldsymbol{K}}, \hat{\boldsymbol{C}} = \underset{\substack{\boldsymbol{K}\geqslant 0, \boldsymbol{C}\geqslant 0 \\ \|\boldsymbol{KC}\|_{\max}\leqslant\Omega}}{\arg\min}\ \Upsilon(\boldsymbol{K}, \boldsymbol{C}) \tag{3.35}$$

Optimization method. Our problem is somewhat similar to the Non-negative Matrix Factorization (NMF) problem, putting the fact that NMF considers a linear model aside. A way to solve the NMF problem is to use Alternating Least Squares (ALS) [PCB10]. We propose to use the same approach for solving Problem (3.35).

The idea of ALS is simple. Instead of solving directly Problem (3.35), we solve two simpler subproblems and alternate the two solutions until convergence. For instance, for estimating $\boldsymbol{K}$, we consider that we have an estimation of $\boldsymbol{C}$, say the estimate $\hat{\boldsymbol{C}}_i$, and we solve the subproblem:

$$\hat{\boldsymbol{K}}_i = \underset{\substack{\boldsymbol{K}\geqslant 0 \\ \|\boldsymbol{K}\hat{\boldsymbol{C}}_i\|_{\max}\leqslant\Omega}}{\arg\min}\ \Upsilon(\boldsymbol{K}, \hat{\boldsymbol{C}}_i) \tag{3.36}$$

Solving Problem (3.36) is easier than solving Problem (3.35) (less parameters) and it gives us an estimation of $\boldsymbol{K}$, the estimate $\hat{\boldsymbol{K}}_i$. We can derive easily a similar subproblem for having a better estimation of $\boldsymbol{C}$ using the estimate $\hat{\boldsymbol{K}}_i$. This will give us a new estimate $\hat{\boldsymbol{C}}_{i+1}$ that we can use for estimating $\boldsymbol{K}$, and so on and so forth. We alternatively solve the two subproblems until the two estimates converge (for instance, by stopping when the cost no longer changes).

This is how ALS works, however, Problem (3.36) is not so straightforward to solve. This is due to the constraint enabling to avoid the saturation area, *i.e.*, $\|\boldsymbol{K}\hat{\boldsymbol{C}}_i\|_{\max} \leqslant \Omega$.

Algorithm 1 Alternating Least Squares for estimating $(\mathbf{K}, \mathbf{C})$

Require: $\hat{\boldsymbol{C}}_0, \Omega, \epsilon, i_{\max}$
  while $|\Upsilon(\hat{\boldsymbol{K}}_{i+1}, \hat{\boldsymbol{C}}_{i+1}) - \Upsilon(\hat{\boldsymbol{K}}_i, \hat{\boldsymbol{C}}_i)| > \epsilon$ & $i < i_{\max}$ **do**
    $\hat{\boldsymbol{K}}_{i+1} = \underset{0 \leqslant \boldsymbol{K} \leqslant \frac{\Omega}{R\|\hat{\boldsymbol{C}}_i\|_{\max}}}{\arg\min} \Upsilon(\boldsymbol{K}, \hat{\boldsymbol{C}}_i)$
    $\hat{\boldsymbol{C}}_{i+1} = \underset{0 \leqslant \boldsymbol{C} \leqslant \frac{\Omega}{R\|\hat{\boldsymbol{K}}_{i+1}\|_{\max}}}{\arg\min} \Upsilon(\hat{\boldsymbol{K}}_{i+1}, \boldsymbol{C})$
    $i = i + 1$
  end while
  return $\hat{\boldsymbol{K}}_i, \hat{\boldsymbol{C}}_i$

To simplify the problem, we propose to relax the constraint by noticing that the $\|\cdot\|_{\max}$ is R-submultiplicative[11]:

$$\|\boldsymbol{K}\hat{\boldsymbol{C}}_i\|_{\max} \leqslant R\|\boldsymbol{K}\|_{\max}\|\hat{\boldsymbol{C}}_i\|_{\max} \tag{3.37}$$

So, instead of upper bounding $\|\boldsymbol{K}\hat{\boldsymbol{C}}_i\|_{\max}$, we can make the problem easier by upper bounding $R\|\boldsymbol{K}\|_{\max}\|\hat{\boldsymbol{C}}_i\|_{\max}$. It is easier in that the influence of the unknown parameter $\boldsymbol{K}$ in $\|\boldsymbol{K}\hat{\boldsymbol{C}}_i\|_{\max}$ is now separated from the influence of the "known" parameter $\hat{\boldsymbol{C}}_i$. So, we can simplify Problem (3.36) by:

$$\hat{\boldsymbol{K}}_i = \underset{0 \leqslant \boldsymbol{K} \leqslant \frac{\Omega}{R\|\hat{\boldsymbol{C}}_i\|_{\max}}}{\arg\min} \Upsilon(\boldsymbol{K}, \hat{\boldsymbol{C}}_i) \tag{3.38}$$

Again, a similar problem than Problem (3.38) can be derived for $\boldsymbol{C}$.

For solving Problem (3.38), we use a quasi-Newton method called L-BFGS [Byr+95] with the `R` package `optim`. This algorithm is particularly suitable here since it can easily incorporate the bound constraints introduced in Problem (3.38) (using the gradient projection method).

The algorithm for estimating $\boldsymbol{K}$ and $\boldsymbol{C}$ is reported in Algorithm 1. In the following, we evaluate this algorithm in simulation.

As a final remark, we noticed in the previous theoretical section that the model is separable either across chemical sensors or across experiments (see Eq. (3.26) & (3.27)). So, Problem (3.38) can be seen as P subproblems. In other words, instead of solving Problem (3.38), it is equivalent to solve the following P subproblems (corresponding to Problem (3.38) for each row):

$$\hat{\boldsymbol{k}}_{p,i}^{\mathrm{T}} = \underset{0 \leqslant \boldsymbol{k}^{\mathrm{T}} \leqslant \frac{\Omega}{R\|\hat{\boldsymbol{C}}_i\|_{\max}}}{\arg\min} \Upsilon(\boldsymbol{k}^{\mathrm{T}}, \hat{\boldsymbol{C}}_i) \tag{3.39}$$

[11] Eq. (3.37) is easy to prove. Let $\boldsymbol{K} \in \mathbb{R}^{P \times R}$ and $\boldsymbol{C} \in \mathbb{R}^{R \times N}$ be two matrices.
Recall that $\|\boldsymbol{K}\boldsymbol{C}\|_{\max} = \max_{ij}(|(\boldsymbol{K}\boldsymbol{C})_{ij}|)$.

$$\begin{aligned} \forall i,j \quad |(\boldsymbol{K}\boldsymbol{C})_{ij}| = |\textstyle\sum_r k_{ir}c_{rj}| &\leqslant \textstyle\sum_r |k_{ir}c_{rj}| \\ &= \textstyle\sum_r |k_{ir}||c_{rj}| \\ &\leqslant \textstyle\sum_r \max_{pq}(|k_{pq}|)\max_{pq}(|c_{pq}|) \\ &= R\|\boldsymbol{K}\|_{\max}\|\boldsymbol{C}\|_{\max} \end{aligned}$$

So, $\|\boldsymbol{K}\boldsymbol{C}\|_{\max} \leqslant R\|\boldsymbol{K}\|_{\max}\|\boldsymbol{C}\|_{\max}$

where $\hat{\boldsymbol{k}}_{p,i}^{\mathrm{T}} \in \mathbb{R}^R$ is the p^{th} row of the i^{th} estimate $\hat{\boldsymbol{K}}_i$. Again, a similar formulation can be easily derived for $\boldsymbol{C}$ by considering the N columns of $\hat{\boldsymbol{C}}_i$. In Algorithm 1, we do not represent the P and N subproblems to keep the formulation compact. However, in the developed R code, Problem (3.38) is solved by considering the P subproblems in Eq. (3.39) (and the same applies to $\boldsymbol{C}$).

3.3.4 Simulation

Simulation settings. To validate Algorithm 1 and assess its performance in the presence of noise, we carry out simulations. The number N of experiments is set to 50. The vector of parameters η_r is $[10, 20, \ldots, 140, 150]^{\mathrm{T}}$ (we take the 2 first parameters when $R = 2$, etc...) and $\Omega = 100$.

We present two scenarios. First, the number of sensors is constant, with $P = 50$, and the number of VOCs in the mixtures is increasing ($R = 2, 5, 10, 15$). Second, the number of VOCs in the mixtures is constant, with $R = 5$, and the number of sensors is increasing ($P = 6, 12, 24, 50$).

We generate the true parameters $\boldsymbol{K}^* \in \mathbb{R}^{P \times R}$ and $\boldsymbol{C}^* \in \mathbb{R}^{R \times N}$ according to an exponential law with $\lambda = 1.5$. The columns of $\boldsymbol{K}^*$ are normalized in order to avoid the saturation regime (and any simulation is discarded if $\|\boldsymbol{K}^*\boldsymbol{C}^*\|_{\max} > \Omega$). The noise is additive and Gaussian with zero-mean and standard deviation σ_n. When negative values are created due to noise addition, we clip them to 0. σ_n depends on the Signal-to-Noise Ratio (SNR) defined as:

$$\text{SNR} = 20\log_{10}\left(\frac{\sigma_s}{\sigma_n}\right) \text{ with } \sigma_s^2 = \frac{\|\boldsymbol{X}\|_F^2}{PN} \tag{3.40}$$

All our simulations are carried out with noise by considering an SNR going from 40dB (little noise) to 0dB (noisy). For each SNR, we carry out 50 different realizations (*i.e.* we generate a new $\boldsymbol{K}^*$ and a new $\boldsymbol{C}^*$, and so a new $\boldsymbol{X}$). For each realization, Algorithm 1 is used, starting from 10 different random initializations $\hat{\boldsymbol{C}}_0$. These 10 initializations may give 10 different solutions. Consequently, for a given realization, we take as final solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{C}})$ the solution having the lowest cost out of the 10. Finally, the criterions used to assess the estimations (see next point) are averaged for a given SNR over the 50 realizations.

Evaluation. As shown in Section 3.3.2, the parameters $(\boldsymbol{K}^*, \boldsymbol{C}^*)$ are identifiable up to a diagonal matrix when the parameter η_r is different for each VOC. In this case, there is theoretically no permutation issue. However, in practice, we observed that permuted versions of $(\boldsymbol{K}^*, \boldsymbol{C}^*)$ are local minima in which Algorithm 1 can converge. In the noise-free case and when R is small, these local minima may not be a problem since their cost is always greater than 0 (the cost taken in $(\boldsymbol{K}^*, \boldsymbol{C}^*)$), so performing several initializations can help to converge to $(\boldsymbol{K}^*, \boldsymbol{C}^*)$. However, there are $R!$ local minima (*i.e.* the number of possible permutations) so even in the noise-free case, when R is large, these local minima are a problem (for instance, if $R = 10$, then there are more than 3 millions of possible local minima). In the noisy case, it may be hard to distinguish between the local minima and $(\boldsymbol{K}^*, \boldsymbol{C}^*)$, whatever the value

of R. Consequently, we have to remove this permutation ambiguity to correctly assess the performance of Algorithm 1.

To remove the permutation ambiguity, we use the correlation between $\hat{\boldsymbol{K}}$ and $\boldsymbol{K}^*$. More precisely, we start by finding the two columns of $\hat{\boldsymbol{K}}$ and $\boldsymbol{K}^*$ which have the highest correlation. These two columns are associated and deleted from the two matrices $\hat{\boldsymbol{K}}$ and $\boldsymbol{K}^*$. The procedure is then repeated for the subsequent columns. The sequence found for rearranging the columns of $\hat{\boldsymbol{K}}$ is also used for rearranging the rows of $\hat{\boldsymbol{C}}$.

To remove the scale ambiguity, we multiply $\hat{\boldsymbol{K}}$ by a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{R\times R}$. The r^{th} diagonal term of $\boldsymbol{D}$ corresponds to $\frac{\|\boldsymbol{k}_r^*\|_2}{\|\hat{\boldsymbol{k}}_r\|_2}$. $\hat{\boldsymbol{C}}$ is multiplied by $\boldsymbol{D}^{-1}$.

As a remark, scale and permutation ambiguities are removed here thanks to the knowledge of $\boldsymbol{K}^*$ (we could have considered equivalently $\boldsymbol{C}^*$). In practice, this is what we are looking for so $\boldsymbol{K}^*$ is unknown. A solution could then be to use some calibrating samples to get rid of these two ambiguities.

To assess the estimates $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{C}})$, we use two measures: a correlation measure[12] and a measure Ψ based on the sum of squared errors:

$$\Psi(\boldsymbol{K}^*, \hat{\boldsymbol{K}}) = \frac{\|\boldsymbol{K}^* - \hat{\boldsymbol{K}}\|_F}{\|\boldsymbol{K}^*\|_F} \quad \text{and} \quad \text{Cor}(\boldsymbol{K}^*, \hat{\boldsymbol{K}}) = \frac{1}{R}\sum_{r=1}^{R} \text{cor}(\boldsymbol{k}_r^*, \hat{\boldsymbol{k}}_r) \tag{3.41}$$

The closer to 0 Ψ, the better. The closer to 1 the correlation measure, the better.

Cost evolution. Convergence properties of Algorithm 1 have not been theoretically investigated yet. In Figure 3.13, we represent the evolution of the cost for 10 different initializations. Algorithm 1 seems to converge, at least to local minima.

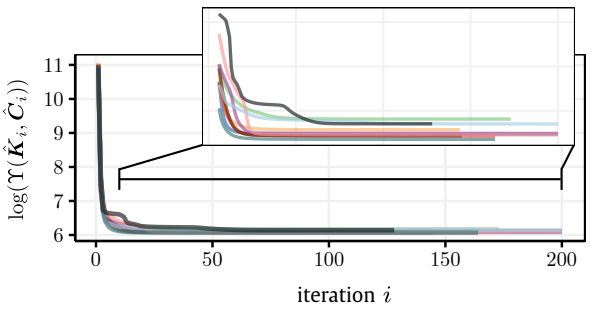


Fig. 3.13.: Evolution of the cost over iterations for 10 random initializations (each color stands for an initilization). Algorithm 1 is stopped if the difference between two successive costs is lower than 10^{-4}, or after 200 iterations. Here, $R = 5, P = N = 50$ and SNR $= 20$dB. Top figure is a zoom.

N & P constants, R variable. In this scenario, the number of VOCs R is increasing while the number N of experiments and the number P of chemical sensors are fixed. The values of N and P are set to 50. Since R is increasing, the number of unknowns increases while the number of equations remains constant. However, even in the hardest case

[12] The correlation is defined as:

$$\forall(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p,\ \text{cor}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

where $\bar{\cdot}$ stands for the mean of the vector.

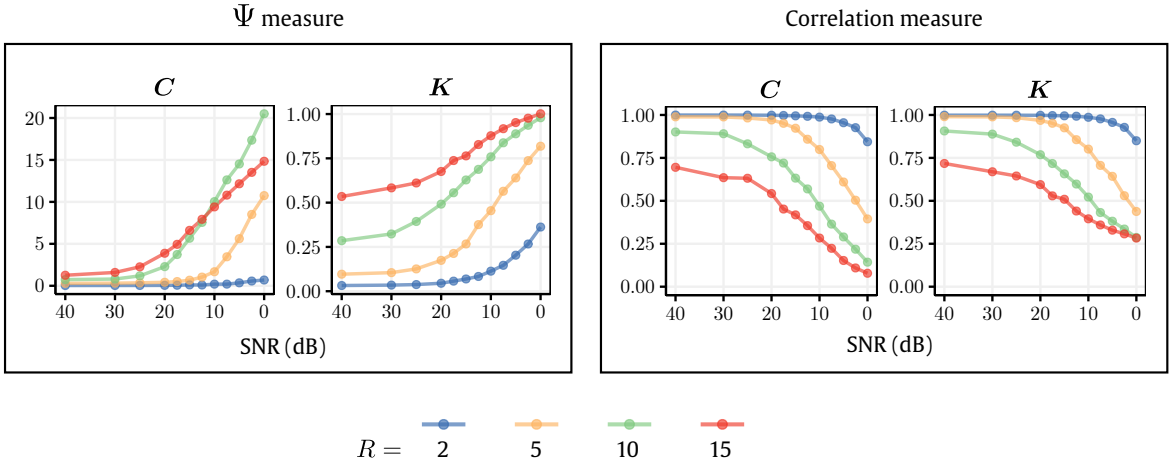


Fig. 3.14.: Results for $N = P = 50$ and $R = 2, 5, 10, 15$. Each color stands for a different R. Left figures correspond to the Ψ measure and right figures to the correlation measure (see Eq. (3.41)). The closer to 0 Ψ, the better. The closer to 1 the correlation measure, the better.

considered here ($R = 15$), the number of equations is much greater than the number of unknowns. To give a general idea of the values of R that we have taken, in the literature, studies with real data often deal with binary gas mixtures and sometimes ternary mixtures. However, to our knowledge, gas mixtures with $R \geqslant 5$ have never been studied in practical applications.

The results are reported in Figure 3.14. For $R = 2$ or $R = 5$, Algorithm 1 performs well when the noise is weak enough. Starting from 20dB, the performance for $R = 5$ rapidly decreases.

The task is much harder for $R = 10$ and $R = 15$, even if the noise is weak, which seems intuitive. When the noise is high, we observed that the upper bound constraint is always active (*i.e.* $\|\hat{\boldsymbol{K}}\|_{\max} = \frac{\Omega}{R\|\hat{\boldsymbol{C}}\|_{\max}}$). This means that Algorithm 1 would likely diverge without the upper bound constraint (but it does not mean that the constraint actually helped the algorithm to converge to a better estimation).

In all the cases, we can notice that Ψ is much higher for $\boldsymbol{C}$ than for $\boldsymbol{K}$ (especially when the noise is high). We explain that by the fact that we always use $\boldsymbol{K}^*$ for estimating a diagonal matrix $\boldsymbol{D}$ enabling to remove the scale ambiguity in $\hat{\boldsymbol{K}}$. $\boldsymbol{D}^{-1}$ is then used for $\hat{\boldsymbol{C}}$. However, when the noise is too high, the estimated parameters $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{C}})$ are then far from the true ones $(\boldsymbol{K}^*, \boldsymbol{C}^*)$. The multiplication by $\boldsymbol{D}$ makes $\boldsymbol{K}^*$ and $\hat{\boldsymbol{K}}$ closer while the multiplication by $\boldsymbol{D}^{-1}$ makes $\boldsymbol{C}^*$ and $\hat{\boldsymbol{C}}$ farther.

As a conclusion, in this scenario, Algorithm 1 performs well if the noise is weak enough. However, what is the SNR value that we should expect with real data? In Section 3.2, we introduced a data set measuring 6 different pure VOCs across 15 experiments each (*i.e.* Data set 1 in Table 3.1). From these measurements, we can extract a data matrix $\boldsymbol{X}$ corresponding to the steady state values (see Figure 3.9, yellow part). From $\boldsymbol{X}$, we can compute the signal part of Eq. (3.40), corresponding to σ_s. We can also compute the noise part by calculating the standard deviation of the response during the baseline acquisition (by taking the 30 first

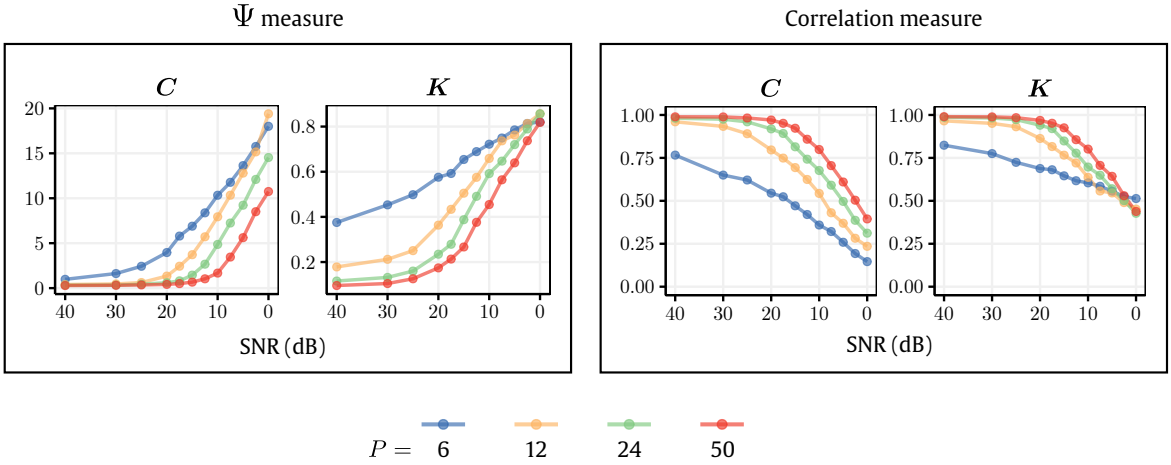


Fig. 3.15.: Results for $N = 50$, $R = 5$ and $P = 6, 12, 24, 50$. Each color stands for a different P. Left figures correspond to the Ψ measure and right figures to the correlation measure (see Eq. (3.41)). The closer to 0 Ψ, the better. The closer to 1 the correlation measure, the better.

time points of a single measurement and by randomly selecting one chemical sensor). If we do so, then we find an SNR value of 42dB, meaning that the noise in real data is rather weak. However, this measure is obviously dependent on several parameters such as the VOCs used or the VOC concentrations (a lower concentration leads to a lower signal). In addition, this value does not take into account the modeling error. The Langmuir-SPRi mixture model is just an approximation of what it is really occurring in real life. The assumptions on which the model relies may not be true leading to another kind of "noise", which is difficult to measure (and we saw in Section 3.2 that some of the assumptions are actually violated in practice).

N & R constants, P variable. In Chapter 1 (Section 1.2.6), we made a short critical review of the literature, pointing out a debate concerning the number of chemical sensors used. We clearly support the side "the more, the better" and this section is appropriate to show the benefits of using more chemical sensors. To that end, the number R of VOCs in the mixtures is now constant ($R = 5$) while the number of sensors is varying ($P = 6, 12, 24, 50$). The idea is then to evaluate the influence of the number of chemical sensors in the Blind Source Separation framework.

The results are reported in Figure 3.15 and they clearly indicate that the more chemical sensors, the better. In fact, for $P = 6$ (a typical value in the literature, see Table 1.1 in Chapter 1), the number of equations ($= 300$) is then quite close to the number of unknowns ($= 280$). To improve the performance, a solution is to increase the number of experiments. However, this solution is time consuming and the value N used here is already high compared to the number of samples actually used in the literature (Section 1.2.6, Chapter 1). Similar results could be obtained by varying N while keeping P and R constants.

3.3.5 Conclusion

In this chapter, we derived the formulation of a theoretical non-linear model in the presence of a gas mixture. The model has been formulated for the equilibrium phase and is based on

the combination of the Langmuir model with the linear approximation of the SPRi response, so we called it the Langmuir-SPRi mixture model.

This model is based on VOCs concentrations, the individual signatures and on some VOCs characteristics. To our knowledge, this model is new so we provided a theoretical analysis concerning the estimation of these parameters for two different problems: the Regression problem (either the concentrations or the signatures are unknown) and the Blind Source Separation problem (both the concentrations and the signatures are unknown). For the latter, we also developed an algorithm in order to estimate both the concentrations and the signatures from a data matrix. This algorithm is based on Alternating Least Squares with bound constraints. We showed that it performs well in simulation when the noise is weak enough and when the number of chemical sensors and experiments is sufficiently high compared to the number of VOCs in the mixtures.

The actual results can be improved in many ways.

Concerning the theoretical part, the identifiability of the Langmuir-SPRi mixture model in the Blind Source Separation framework has not been proven. In fact, this question is remarkably difficult even if one considers the same problem but with a linear model. This problem is referred to as Non-negative Matrix Factorization (NMF) and represents a good example of the difficulty of the question. In fact, many works have been carried out during the past two decades in order to find conditions under which NMF becomes identifiable (see for instance [DS04; Lau+08; HSS14; FHS18]). In the near future, we may take inspiration from these works in order to find identifiability conditions for the Langmuir-SPRi mixture model (*e.g.* we could assume the existence of "pure" mixtures in $\boldsymbol{X}$).

NMF is actually useful in many practical applications (for instance with spectroscopy measurements [Mou+06]). However, we did not prove yet if the Langmuir-SPRi mixture model is useful. In Chapter 6, we will see that a linear model is in fact already sufficient to provide good results with real data obtained in robotic applications. So, a crucial next step is to validate the Langmuir-SPRi mixture model on real data and to verify in what extent it improves the results compared to a linear model. To that end, experiments in a controlled environment with a reliable measure of the VOC concentrations must be conducted (for instance by taking inspiration from [Mad+18]).

Concerning the algorithmic part, we made the choice of Alternating Least Squares but other algorithms can be obviously tested. To improve the solution, we could also consider additional constraints such as sparsity (this constraint will be used in Chapter 6 by considering an ℓ_1-penalization).

4 Drift

Contents

"[...] an indescribable chaos of odours reigned in the House of Baldini. However exquisite the quality of individual items - for Baldini bought wares of only highest quality - the blend of odours was almost unbearable, as if each musician in a thousand-member orchestra were playing a different melody at fortissimo. Baldini and his assistants were themselves inured to this chaos, like ageing orchestra conductors (all of whom are hard of hearing, of course)"

Patrick Süskind, *Perfume, the Story of a Murderer*[1].

Baldini is an old perfurmer who teaches to Grenouille the foundations of perfumery. By being surrounded all his life by thousands of different odors, the ability of his nose is now deteriorated by its long years of use. In fact, his nose grew older and does no longer perceive the scents in the same way than in his youth. Ageing affects also chemical sensors. It is one of the main factors causing an unexpected phenomenon: the classifier will not generalize to future samples if too much time has elapsed since the training set was acquired. This issue is called drift and all this chapter is devoted to it.

The chapter is organized as follows. First, we detail the drift issue: its causes and its consequences. Then, we describe state-of-the-art methods to compensate for the drift. Third, we introduce two new methods which do not require any label from the other measurement days. Then, the proposed methods are evaluated on an artificial data set. Finally, we introduce a data set acquired over 9 months, suffering from drift. All the introduced methods are then compared, based on this data set.

4.1 Introduction

What is "drift"?

Drift is a dynamical process occurring over time and affecting the reproducibility of the instrument [Hol+97]. Drift is responsible for the lack of stability of chemical sensors and is considered as one of the main factors explaining the low number of practical applications with these sensors [Pad+10].

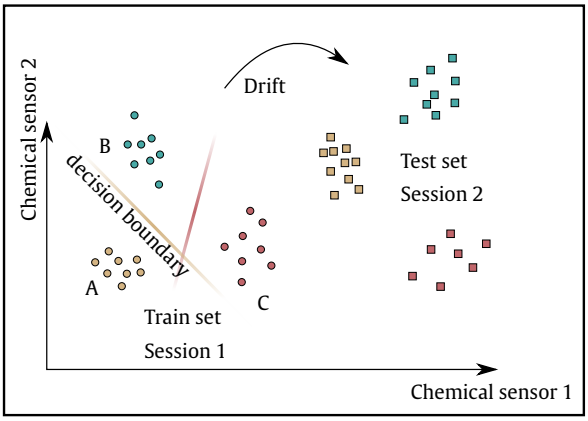


Fig. 4.1.: Simplified representation of the drift issue. Here, during Session 2, all new points are classified as C due to the drift.

Let us take an example to make the consequences of the drift more concrete. Let us assume that we want to build an application targeting the classification of 3 differents VOCs, say A, B and C. During the first measurement session (called Session 1), we start by realizing some time-consuming experiments during which we measure several times these target VOCs. From these labelled

[1]Translated from the German by John E. Woods.

measurements, we extract the signatures of the VOCs and build a data base. From this data base, a classifier learns to differentiate between the target VOCs A, B and C. Then, it comes time for using the eNose in the field, so we carry out new experiments during another measurement session (Session 2) occurring during another day.

Unexpectedly, all the new measurements are classified as C, even if A and B are also present. So, what happened? Drift. Drift occurs between the two sessions, meaning that the signatures learnt from Session 1 are not the same as the signatures acquired during Session 2. Consequently, the classifier becomes useless and needs to be retrained, or the signatures from Session 2 need to be corrected. A simplified representation of the drift issue is reported in Figure 4.1.

What are the causes of the drift?

The causes of the drift are numerous and depend on the sensing and transduction mechanisms [DCF12]. One generally splits the drift into two parts depending on the causes [HGO08]. First, the drift related to the chemical sensors themselves and second, the drift of the measurement system.

Drift related to the chemical sensors themselves corresponds to chemical and physical changes of sensing materials. Two main causes are often stressed. The first cause is sensor poisoning. Sensor poisoning is due to the injection of some VOCs which have led to irreversible reactions with the sensing materials. These reactions result in the blocking or in the creation of reaction sites [HA04]. In our case and in the case of semiconductor sensors [HTK00], sulphur compounds and acids can lead to such reactions. The second cause is sensor ageing. Sensor ageing can be due to the spontaneous re-organization of the sensing material [HA04], leading to a change in the number of reaction sites. These sources of drift will generally impair the sensitivity of the instrument for the target VOCs.

The drift related to the measurement system itself can be due to changes in temperature, humidity or pressure [HGO08]. These external parameters will influence the chemical reactions, and this also applies to the signatures. Memory effects can also take place. They are due to some traces of VOCs remaining in the sampling system or on the sensing surface. To avoid memory effects, a small size sampling system[2] and a long cleaning phase are desirable [HA04]. Third, a short term drift spreading generally over several minutes or hours is often observed. This can be due to the heating of the electronic system or due to the instrument warm-up (a short period of time during which the response of chemical sensors can be unstable) [HA04]. Finally, a chemical drift of the VOCs themselves can appear due for instance to a bad storage of the samples (*e.g.* oxidation). If the final goal of the application is

[2]It happens that VOCs not only interact with the eNose but also with the sampling system. For instance, VOCs can adsorb on the side of the pipes (and then desorb after an unknown time). This undesirable reaction can disturb the next measurement with another VOC if the remaining VOC desorbs from the sampling system at the same time. To prevent this issue, the sampling system must be small to limit the interaction and ideally made with materials which are little prone to interact with VOCs (*e.g* PolyEtherEtherKetone tubing).

Type		Causes (examples)
Measurement system	Memory effects	Too short cleaning phase
	Chemical drift	Bad storage of the samples
		Change in concentration (if classification task)
	Short-term drift	Sensor warm-up, electronic system heating
	Environmental parameters	Changes in temperature, humidity, pressure
Sensor drift	Sensor poisoning	Irreversible binding reactions
	Sensor ageing	Re-organization of the sensor surface

Tab. 4.1.: Summary of some causes of the intra-instrument drift [HGO08; HA04].

only VOC classification (and not VOC quantification), then variations in concentration can also be considered as a chemical drift. Some components of the drift related to the measurement system can be avoided when the experiments are carried out in a controlled environment (see for instance the study of Vergara et al. [Ver+12]). However, in the field, the sampling cannot be always reliably controlled. The possible sources of the drift are summarized in Table 4.1.

Finally, all the causes previously introduced impair generalization performance of a single instrument: we call this drift, the **intra-instrument drift**. However, the final goal is to use one or several eNoses to generate a data base for training a classifier, and then generalize to other eNoses deployed in the field. In this case, another drift will appear, an **inter-instrument drift**. In fact, the manufacturing process of chemical sensors is rarely reproducible [RN10]. This means that the classifier which learns from an eNose 1 may not directly generalize to an eNose 2. In all this chapter, we focus only on the intra-instrument drift.

What are the solutions?

A naive solution is to briefly retrain the classifier before testing the eNose in the field. Ideally, the regeneration of the training samples is also performed in the same conditions (*e.g.* temperature, humidity) than the test samples. Obviously, this solution is not desirable at all since each experiment is time-consuming.

Therefore, the solution adopted in the literature is to design methods compensating for the drift. Ideally, these methods would correct individually each component of the drift (ageing, temperature, humidity, etc...). However, the main difficulty is that all the causes of the drift can be concomitant in realistic experiments, so their effects may not be separable. Consequently, all these causes are often treated as a single drift.

In the following, we review the main correction methods. All the methods described require a **Fundamental Hypothesis**: all the individual drifts of each target VOC must be related in some way (in the simplest case, correlated). In other words, if each class drifts along a random direction which is dependent on each target VOC, then there is nothing to do and all

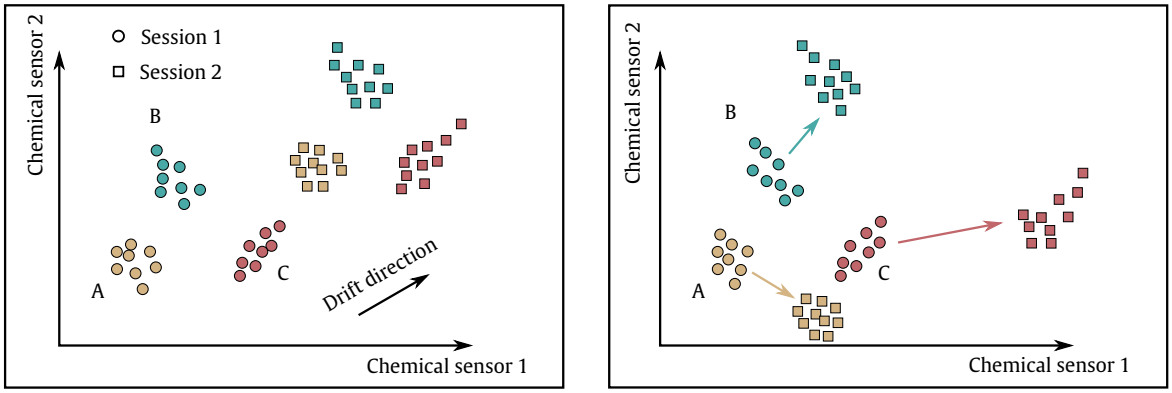


Fig. 4.2.: Simplified representation of the Fundamental Hypothesis. Left figure: Fundamental Hypothesis is valid meaning that all the classes drift along a common direction. Right figure: Fundamental Hypothesis is violated meaning that each class drifts along a random direction.

the following methods will fail. A simplified representation of the Fundamental Hypothesis is represented in Figure 4.2.

4.2 State of the art methods

The literature on drift correction methods is prolific and we decide to review only the main contributions. We especially focus on drift correction methods which have been successfully applied on unseen real data (not only simulated data) with several sessions sufficiently spaced in time (not only real data with 1 week of recordings). Ideally, these methods have been tested by different authors. Besides, correction methods that we discuss can be applied in our case: concentration is neither controlled nor measured, environmental conditions may vary, and the final goal is VOC classification. A more comprehensive review of the existing techniques can be found in [MGG12] or in [Rud18].

First, drift is often seen as a **gradual change** of the chemical sensors [Art+00; Ziy+10] or as a **slow-varying process** over time [Pad+10; DC+11]. This means that if a measurement session lasts for only few hours, the drift affecting the samples in this time window is assumed to be constant. In other words, the drift occurs between measurement sessions but not inside a measurement session (lasting for few hours).

Second, we can differentiate between two kinds of drift which can be present in the data: we call them **Continuous drift** and **Discontinuous drift**. Continuous drift assumes that little time has elapsed between the measurement sessions (*e.g.* one session per day) or that the environmental conditions are controlled. In this case, the drift between two consecutive sessions is small enough to be considered as a "continuous" change of the chemical sensors. At the other end, it happens that a long time elapses between two consecutive sessions or that a large change of environmental conditions occurs (especially in practical applications). In this case, a large drift occurs between two consecutive sessions, causing a Discontinuous drift. A representation of Continuous drift and Discontinuous drift can be found in Figure 4.3. All the methods in the literature are not able to cope with these two types of drift. In

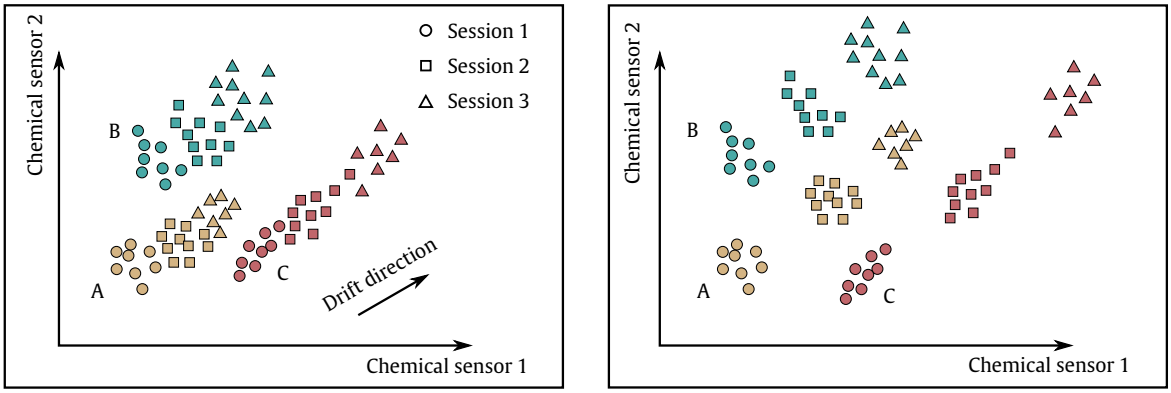


Fig. 4.3.: Simplified representation of the notion of Continuous and Discontinuous Drift. Left figure: the drift between sessions is small enough to be considered as a gradual change (Continuous drift). Right figure: the drift between sessions is large (Discontinuous drift).

this chapter, we introduce a data set with Discontinuous drift and, in Chapter 5 we present another data set suffering from a Continuous drift.

Then, we can identify three different groups of methods which can be classified according to the required assumptions on the labelled data available for drift correction.

First, a classical approach is to assume the existence of calibrating samples of one or several reference VOCs during each measurement session. Reference VOCs must have a correlated drift with the drift of the target VOCs [HTK00]. The calibrating samples are then used to estimate the drift directions. We call this scenario: **Calibrant Scenario** (CS). Hereinafter, we detail PCA-CC and PLS-CC from Artursson et al. [Art+00]. Both methods can cope with Continuous drift and Discontinuous drift.

Second, another approach relaxes the assumption of calibrating samples. In fact, finding reference VOCs is not an easy task and requires a long upstream work which must be repeated anytime an application with new target VOCs must be designed. In this second group of methods, the idea is rather to assume that the drift has been captured by training data. In a classification task, a training set is always required and can be acquired during a single measurement session. Here, the authors assume that the training set has been acquired across multiple sessions over time. If a long enough time has elapsed between sessions, then the training set must contain drift information. So this group of methods extracts the drift directions directly from the training set and assumes that these directions will no longer evolve for subsequent sessions. We call this scenario: **Multi-Session Scenario** (MSS). Hereinafter, we detail OSC from Padilla et al. [Pad+10] and CPCA-CC from Ziyatdinov et al. [Ziy+10]. Both methods can cope with Continuous drift and Discontinuous drift.

Finally, the last group of methods assumes that the only labelled data available is the training set, from a single measurement session (so drift is not present). In fact, MSS can be attractive since it does not require any reference VOC but it requires a long experimental time. For instance, the results of OSC [Pad+10] were obtained with data acquired over 15 days (the

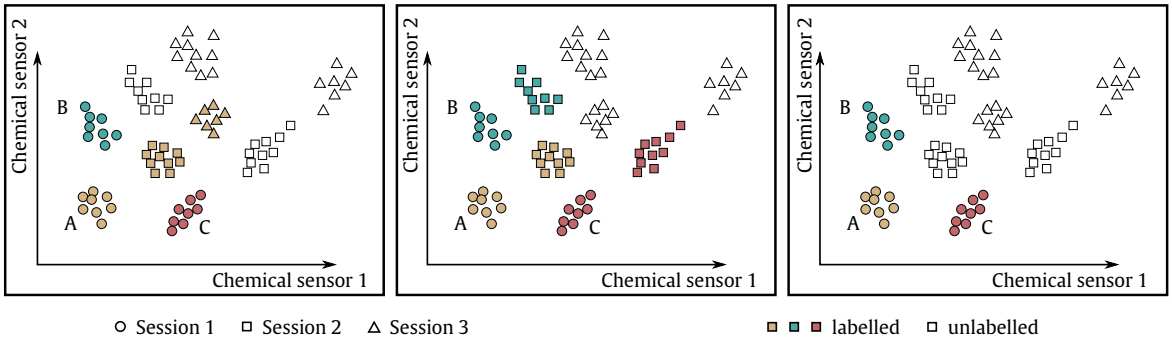


Fig. 4.4.: Simplified representation of the different scenarios that we proposed to classify the different state-of-the-art methods. Left figure: calibrating samples are measured during each session (Calibrant Scenario), here A is the calibrant. Middle figure: the training set includes several measurement sessions over time, and we assume that it contains drift information (Multi-Session Scenario). Right figure: the training set includes only Session 1 and there is no other labelled samples (Blind Scenario). We represent a Discontinuous drift but the same applies to a Continuous drift.

number of measurement sessions is not detailed). For the results of CPCA-CC, Ziyatdinov et al. used at least one month of recordings [Ziy+10]. This long experimental time is limiting so other authors try another more appealing approach: they simply assume that there is no labelled data for estimating the drift. So the directions of the drift between two different sessions must be estimated blindly. We call this scenario: **Blind Scenario** (BS). Only few methods have been reported in the literature for dealing with the BS and they often rely on a classification step. The classification step identifies the samples and helps therefore to estimate the drift direction. So these methods can be generally applied only for Continuous drift but will fail in the presence of Discontinuous drift. Hereinafter, we detail the evolutionary-based method of Di Carlo et al. [DC+11].

All these scenarios clearly differ in difficulty, even if all of them are challenging. By increasing order of difficulty, we can rank the scenarios as: CS, MSS and BS. The three scenarios are represented in Figure 4.4.

In the following sections, without loss of generality, we assume that we have two different sessions. We note $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}$ the data acquired from Session 1. This first measurement session is entirely labelled (with labels $\boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}$) and will be later used as the training set. We note $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ another session, affected by some drift compared to Session 1.

4.2.1 Calibrant Scenario

The first group of methods assumes the existence of calibrating samples acquired during each session. So, for each session, the samples from some reference VOCs are labelled. We note $\boldsymbol{X}_{\text{cal}}$ the matrix containing all these calibrating samples (in all the different sessions). The two methods described here are based on a technique called Component Correction (CC). The main assumption behind Component Correction is that the drift follows a preferred direction which can be removed by subtraction. In the Calibrant Scenario, this direction can be estimated with the calibrating samples. The way this direction is estimated then depends on the method used.

Algorithm 2 PCA-CC from Artursson et al. [Art+00]

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, \boldsymbol{X} \in \mathbb{R}^{N \times P}, \boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}, \boldsymbol{\ell}_{\text{cal}}, K \in \mathbb{N}$
Build $\boldsymbol{X}_{\text{cal}}$
$\boldsymbol{T}, \boldsymbol{P} = \text{PCA}(\boldsymbol{X}_{\text{cal}})$ # $\boldsymbol{X}_{cal} = \boldsymbol{T}\boldsymbol{P}^T$
$\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{P}_K\boldsymbol{P}_K^{\text{T}})$ # $\boldsymbol{P}_K \in \mathbb{R}^{P \times K}$ *the* K *first columns of* $\boldsymbol{P}$

Algorithm 3 PLS-CC from Artursson et al. [Art+00]

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, \boldsymbol{X} \in \mathbb{R}^{N \times P}, \boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}, \boldsymbol{\ell}_{\text{cal}}, K \in \mathbb{N}$
Build $\boldsymbol{X}_{\text{cal}}, \boldsymbol{z}_{\text{cal}}$
$\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{T}, \boldsymbol{C}, \boldsymbol{U} = \text{PLS}(\boldsymbol{X}_{\text{cal}}, \boldsymbol{z}_{\text{cal}}, K)$ # *see [CBC21] for details about PLS*
$\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{W}_K(\boldsymbol{P}_K^{\text{T}}\boldsymbol{W}_K)^{-1}\boldsymbol{P}_K^{\text{T}})$ # $\boldsymbol{W}_K$ *the* K *first columns of* $\boldsymbol{W}$

The most used method is probably the one of Artursson et al. [Art+00], called PCA-CC. In this case, the drift direction is estimated with the help of Principal Component Analysis (PCA). The idea is quite simple: if enough drift occurred between the sessions, then most of the variance in $\boldsymbol{X}_{\text{cal}}$ can be attributed to the drift. An easy way to compute the direction having the largest variance is then to compute the first principal direction $\boldsymbol{p}_1 \in \mathbb{R}^P$ of the calibrating matrix $\boldsymbol{X}_{\text{cal}}$. Afterwards, one simply performs an orthogonal projection of $\boldsymbol{X}$ against $\boldsymbol{p}_1$ to remove all the information carrying by the direction $\boldsymbol{p}_1$ (for which $\boldsymbol{p}_1^{\text{T}}\boldsymbol{p}_1 = 1$) in $\boldsymbol{X}$: $\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{p}_1\boldsymbol{p}_1^{\text{T}})$. This procedure can be repeated several times, say K times, with the next principal directions of $\boldsymbol{X}_{\text{cal}}$ if the drift follows several directions. PCA-CC is described in Algorithm 2.

In the same article, Artursson et al. proposed an alternative version of PCA-CC, called PLS-CC [Art+00]. Instead of estimating the drift direction as the principal directions of $\boldsymbol{X}_{\text{cal}}$, they used PLS (Partial Least Squares [WSE01]). PLS is a quite used technique in Chemometrics in order to describe some dependent variables $\boldsymbol{Z}$ as a linear combination of predictors $\boldsymbol{R}$. In fact, PLS does more than a simple linear regression. PLS estimates some components which explain both $\boldsymbol{R}$ and $\boldsymbol{Z}$ while maximizing the variance of the cross-product $\boldsymbol{Z}^{\text{T}}\boldsymbol{R}$ (see [CBC21] for details about the PLS method). Artursson et al. proposed to use measurement time as the dependent variable $\boldsymbol{z}_{\text{cal}}$ since drift is time-related and $\boldsymbol{X}_{\text{cal}}$ as the predictors. Then, they correct $\boldsymbol{X}$ by using the first latent vectors of $\boldsymbol{X}_{\text{cal}}$: $\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \frac{\boldsymbol{w}_1\boldsymbol{p}_1^{\text{T}}}{\boldsymbol{w}_1^{\text{T}}\boldsymbol{p}_1})$. If the drift follows several directions, several latent components can be removed in the same way. PLS-CC is described in Algorithm 3.

The methods in the Calibrant Scenario will likely cope with Continuous and Discontinuous drifts if all the drift directions are captured by the calibrating samples.

4.2.2 Multi-Session Scenario

The second group of methods assumes that several sessions have been labelled, they correspond to the first measurement sessions, say the first S_t sessions, and correspond to several weeks of recordings. We note $\boldsymbol{X}_{1:S_t} \in \mathbb{R}^{(N_1+\ldots+N_{S_t}) \times P}$ and $\boldsymbol{L}_{1:S_t} \in \mathbb{R}^{(N_1+\ldots+N_{S_t}) \times R}$ respectively the data matrix containing the first S training sessions and the class-membership matrix

Algorithm 4 CPCA-CC from Ziyatdinov et al. [Ziy+10]

Input: $\boldsymbol{X}_{1:S_t} \in \mathbb{R}^{(N_1+...+N_{S_t})\times P}$, $\boldsymbol{L}_{1:S_t} \in \mathbb{R}^{(N_1+...+N_{S_t})\times R}$, $\boldsymbol{X} \in \mathbb{R}^{N\times P}$, $K \in \mathbb{N}$
Compute the covariance matrices $\boldsymbol{S}_r$ from $\boldsymbol{X}_{1:S_t}$
$\boldsymbol{V}, \{\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_R\} = \text{JD}(\boldsymbol{S}_1, ..., \boldsymbol{S}_R)$ *# Joint Diagonalization (from [CS96])*
$\text{SSE}_p = \sum_{r=1}^{R} \|\boldsymbol{S}_r - \sigma_{pp,r}\boldsymbol{v_p}\boldsymbol{v_p}^{\text{T}}\|_F^2$
 # Note $\boldsymbol{P}_K$ *the* K *first columns of the reordered* $\boldsymbol{V}$ *(by increasing SSE)*
$\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{P}_K\boldsymbol{P}_K^{\text{T}})$
$\boldsymbol{X}_{1:S_t,\text{cor}} = \boldsymbol{X}_{1:S_t}(\boldsymbol{I}_P - \boldsymbol{P}_K\boldsymbol{P}_K^{\text{T}})$ *# Recall that* $\boldsymbol{X}_{1:S_t}$ *contains also drift*

containing 0s and 1s. Each row $\boldsymbol{\ell}_n^{\text{T}} \in \mathbb{R}^R$ of $\boldsymbol{L}_{1:S_t}$ contains a single 1 indicating the class membership (if the r^{th} element of $\boldsymbol{\ell}_n^{\text{T}}$ is 1, then the sample n belongs to the class of VOC r).

A calibrant-free version of the PCA-CC of Artursson et al. [Art+00] has been later proposed by Ziyatdinov et al. [Ziy+10], called CPCA-CC. They proposed to use Common Principal Component Analysis (CPCA) from B. N. Flury [Flu84]. Given the sample covariance matrices $\boldsymbol{S}_r \in \mathbb{R}^{P\times P}$ of each class, CPCA aims at finding an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{P\times P}$ able of a simultaneous diagonalization of the R sample covariance matrices, $\boldsymbol{\Sigma}_r = \boldsymbol{V}^{\text{T}}\boldsymbol{S}_r\boldsymbol{V}$ with $\boldsymbol{\Sigma}_r$ a diagonal matrix. Mathematically, it assumes that the eigendecomposition of each sample covariance matrix can be performed with a common set of eigenvectors and different eigenvalues. To understand the meaning of these eigenvectors for $\boldsymbol{X}_{1:S_t}$, it helps to consider only one class. In fact, with only one class, CPCA is equivalent to a simple PCA so the first eigenvector indicates the direction of maximum variance of the class. When multiple classes are present, the first eigenvector provided by CPCA will no longer reflect the direction of maximum variance of a single class but the direction of variance common for all classes. In the case of drift, this variance is assumed to be drift-related since the drift will influence all the classes. The directions estimated by the CPCA of $\boldsymbol{X}_{1:S_t}$ are then used to correct the drift in $\boldsymbol{X}$ by orthogonal projection. Since $\boldsymbol{X}_{1:S_t}$ includes some drift, we correct it in the same way. CPCA is known as Joint Diagonalization (JD) in the Signal Processing community. Ziyatdinov et al. used the algorithm of Cardoso and Souloumiac [CS96] for performing the JD of the sample covariance matrices.

Regarding the matrices $\boldsymbol{\Sigma}_r$ provided by CPCA, some precautions must be taken when one performs Joint Diagonalization or CPCA. In fact, compared to PCA in which eigenvalues are simply ranked by decreasing order, most algorithms for JD will rank the diagonal values $\sigma_{pp,r}$ of $\boldsymbol{\Sigma}_r$ in an arbitrary order. Thus, the first eigenvector may not be the direction of maximum common variance. Ziyatdinov et al. [Ziy+10] proposed to choose as first eigenvector the one which has the greatest *Signal-to-Noise Ratio* (SNR). However, this SNR value is not detailed in their paper. Here, we propose to use the following metric:

$$\text{SSE}_p = \sum_{r=1}^{R} \|\boldsymbol{S}_r - \sigma_{pp,r}\boldsymbol{v_p}\boldsymbol{v_p}^{\text{T}}\|_F^2$$

with $\|\cdot\|_F$ the Frobenius norm (sum of the squared terms of a matrix).

Algorithm 5 OSC from Padilla et al. [Pad+10]

Input: $\boldsymbol{X}_{1:S_t} \in \mathbb{R}^{(N_1+\ldots+N_{S_t})\times P}, \boldsymbol{L}_{1:S_t} \in \mathbb{R}^{(N_1+\ldots+N_{S_t})\times R}, \boldsymbol{X} \in \mathbb{R}^{N\times P}, K \in \mathbb{N}$
$\boldsymbol{W}, \boldsymbol{P} = \text{OSC}(\boldsymbol{X}_{1:S_t}, \boldsymbol{L}_{1:S_t}, K)$ *# OSC based on Direct Orthogonalization from [And99]*
$\boldsymbol{X}_{\text{cor}} = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{P}\boldsymbol{P}^{\text{T}})$ *#* $\boldsymbol{P} \in \mathbb{R}^{P\times K}$
$\boldsymbol{X}_{1:S_t,\text{cor}} = \boldsymbol{X}_{1:S_t}(\boldsymbol{I}_P - \boldsymbol{P}\boldsymbol{P}^{\text{T}})$ *# Recall that* $\boldsymbol{X}_{1:S_t}$ *contains also drift*

The lower the SSE is, the better the set $(\sigma_{pp,r}, \boldsymbol{v_p})$ explains the covariance matrices $\boldsymbol{S}_r$. So, the eigenvectors are then ordered by increasing SSE. CPCA-CC is described in Algorithm 4.

Padilla et al. [Pad+10] proposed to use Orthogonal Signal Correction (OSC), a method which has been successfully applied to the correction of near infrared spectra [Wol+98]. OSC finds some directions explaining variance in $\boldsymbol{X}_{1:S_t}$ while being orthogonal to $\boldsymbol{L}_{1:S_t}$. The idea is then to remove these directions unrelated to $\boldsymbol{L}_{1:S_t}$ which are assumed to be drift-related. Many different algorithms have been proposed for OSC (see [WJS01]), Padilla et al. [Pad+10] used the MATLAB implementation from Wise and Gallagher [WG]. Since we do not have access to the MATLAB code, we prefer to use another OSC algorithm proposed by C. A. Andersson [And99], called Direct Orthogonalization. This algorithm orthogonalizes $\boldsymbol{X}_{1:S_t}$ against $\boldsymbol{L}_{1:S_t}$ and computes the Principal directions $\boldsymbol{P}$ of the new data matrix. The method of Padilla et al. [Pad+10] is reported in Algorithm 5.

The methods considering the Multi-Session Scenario will likely cope with Continuous and Discontinuous drift if the drift directions do not evolve for further sessions.

4.2.3 Blind Scenario

The last group of methods is the least restrictive in practice. It assumes that there is no labelled data for estimating the drift, so the only labelled data come from the training set acquired during the first measurement session. While being the more realistic, this group of methods has received little attention in the field and is clearly the most challenging. In the following, only $\boldsymbol{X}_1 \in \mathbb{R}^{N_1\times P}$ is labelled (we note $\boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}$ the labels).

Di Carlo et al. [DC+11] proposed an evolutionary-based method to compensate for the drift in the BS. The algorithm proceeds by updating a correction matrix from one session to another, so we consider several measurement sessions $\boldsymbol{X}_2 \in \mathbb{R}^{N_2\times P}, \ldots, \boldsymbol{X}_S \in \mathbb{R}^{N_S\times P}$ which are unlabelled. In fact, the term evolutionary-based method is a bit misleading since the authors simply assume a linear model between the matrix $\boldsymbol{X}_i$ and $\boldsymbol{X}_1$. If $N_1 = N_i$ and the sampling order is the same for Session 1 and Session i, then the method assumes the following model: $\boldsymbol{X}_1 \approx \boldsymbol{X}_i\boldsymbol{M}_i$ (in their article $\boldsymbol{M}_i$ is replaced by $\boldsymbol{I}_P + \boldsymbol{M}_i$). The "evolutionary" part of their algorithm comes from the way they estimate $\boldsymbol{M}_i$: they used an optimization method called Covariance Matrix Adaptation Evolution Strategy (CMA-ES) from N. Hansen and A. Ostermeier [HO96]. The approach of Di Carlo et al. proceeds in several stages. Initially, a classifier learns from $\boldsymbol{X}_1$ (here we use SVM but other classifiers can be chosen) and $\hat{\boldsymbol{M}}_1 = \boldsymbol{I}_P$. Then, samples from $\boldsymbol{X}_2$ are corrected starting from $\hat{\boldsymbol{M}}_2 = \hat{\boldsymbol{M}}_1$, $\boldsymbol{X}_{2,\text{cor}} = \boldsymbol{X}_2\hat{\boldsymbol{M}}_2$. Afterwards, the samples of $\boldsymbol{X}_{2,\text{cor}}$ are classified, leading to a class vector $\boldsymbol{\ell}_2 \in \mathbb{R}^{N_2}$. Based on the prediction

Algorithm 6 Evolutionary-based method from Di Carlo et al. [DC+11]

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, ..., \boldsymbol{X}_S \in \mathbb{R}^{N_S \times P}, \boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}$
SVMtrained = SVM($\boldsymbol{X}_1, \boldsymbol{\ell}_1$) *# other classifiers can be used*
Compute the R covariance matrices $\boldsymbol{S}_r$ from $\boldsymbol{X}_1$
$\hat{\boldsymbol{M}}_1 = \boldsymbol{I}_P$
for i in $2 : S$ **do**
  $\hat{\boldsymbol{M}}_i = \hat{\boldsymbol{M}}_{i-1}$
  $\boldsymbol{X}_{i,\text{cor}} = \boldsymbol{X}_i \hat{\boldsymbol{M}}_i$
  $\boldsymbol{\ell}_i$ = predict(SVMtrained, $\boldsymbol{X}_{i,\text{cor}}$)
  $\hat{\boldsymbol{M}}_i = \arg\min_{\boldsymbol{M}} \sum_{n=1}^{N_i} (\boldsymbol{M}^{\mathrm{T}} \boldsymbol{x}_{i,n} - \boldsymbol{\mu}_{\ell_{i,n}})^{\mathrm{T}} \boldsymbol{S}_{\ell_{i,n}}^{-1} (\boldsymbol{M}^{\mathrm{T}} \boldsymbol{x}_{i,n} - \boldsymbol{\mu}_{\ell_{i,n}})$
  $\boldsymbol{X}_{i,\text{cor}} = \boldsymbol{X}_i \hat{\boldsymbol{M}}_i$
  $\boldsymbol{\ell}_i$ = predict(SVMtrained, $\boldsymbol{X}_{i,\text{cor}}$)
end for

$\ell_{2,n}$ of each sample n, $\hat{\boldsymbol{M}}_2$ is updated by minimizing the following cost function (here, we used Mahalanobis distance but, again, other distance functions can be used):

$$\hat{\boldsymbol{M}}_2 = \arg\min_{\boldsymbol{M}} \sum_{n=1}^{N_2} (\boldsymbol{M}^{\mathrm{T}} \boldsymbol{x}_{2,n} - \boldsymbol{\mu}_{\ell_{2,n}})^{\mathrm{T}} \boldsymbol{S}_{\ell_{2,n}}^{-1} (\boldsymbol{M}^{\mathrm{T}} \boldsymbol{x}_{2,n} - \boldsymbol{\mu}_{\ell_{2,n}})$$

with $\mu_{\ell_{2,n}}$ and $\boldsymbol{S}_{\ell_{2,n}}$ the centroid and the covariance matrix of the class $\ell_{2,n}$ in Session 1.

Once $\hat{\boldsymbol{M}}_2$ has been estimated with CMA-ES, $\boldsymbol{X}_2$ is corrected with the new $\hat{\boldsymbol{M}}_2$ and corrected samples $\boldsymbol{X}_{2,\text{cor}} = \boldsymbol{X}_2 \hat{\boldsymbol{M}}_2$ are again classified. Finally, for Session 3, the process is repeated starting from $\hat{\boldsymbol{M}}_3 = \hat{\boldsymbol{M}}_2$. The method is summarized in Algorithm 6.

It is easy to see that the method of Di Carlo et al. will, in the best case, only cope with Continuous drift. In fact, if two consecutive sessions $(\boldsymbol{X}_{i-1}, \boldsymbol{X}_i)$ suffer from a large drift (*i.e.* Discontinuous drift), then the samples $\boldsymbol{X}_i$ will likely be misclassified by the first classification step since $\boldsymbol{M}_{i-1}$ will not be a good estimator of $\boldsymbol{M}_i$. Even after the optimization process (CMA-ES), $\boldsymbol{M}_i$ will not be correctly estimated because the optimization process will be based on these misclassifications (which are rigid in the optimization). Besides, this wrong estimation will affect the next corrections since their method is incremental. So, in case of Discontinuous drift, one will have to restart the training phase from $\boldsymbol{M}_i$.

Another series of works is based on clustering algorithms and especially on Self-Organized Maps (SOM) ([NDD95; Mar+98; DSP02; Zup+04]) from Kohonen [Koh82]. SOM is an unsupervised artificial neural network which can be used to map data into a lower dimensional space (usually, 2). It can also be used for clustering since SOM tends to preserve the initial distances in the high dimensional space. For VOC classification, a supervised step is however required during the training to identify each class in the map created by SOM. The idea of using such algorithms is to keep training them during the testing phase (Session 2, 3, etc...), so the map can adapt to new samples which are drifting. However, if a large drift occurred, then the readaptation will likely fail.

4.2.4 Summary

The methods considering the Calibrant Scenario (CS) or the Multi-Session Scenario (MSS) will work well in practice as long as the assumptions behind the CS (the calibrant has the same drift than the target VOCs) and the MSS (after few sessions, drift direction no longer changes) remain valid. They will work for both Continuous and Discontinuous drift.

The methods considering the Blind Scenario (BS) are much more challenging and often rely on a classification step, useful for estimating the drift direction. These methods may work in the presence of Continuous drift but will likely fail in the presence of Discontinuous drift.

As a final note on this state of the art, I have implemented all the algorithms previously described in R. The implementations of Algorithm 2, 3, 4, 5 and 6 and some simulated data can be obtained at:

`https://github.com/mahopie/eNoseDrift.git`

4.3 A correction method in the Blind Scenario

The methods which consider the Blind Scenario generally rely on either the updating of clustering algorithms or the definition of a drift model. In the following, we derive the mathematical formulation of a drift model.

In an ideal word, the signature $\tilde{\boldsymbol{x}}_r$ of a VOC r at a concentration c_r is deterministic by nature. It is due to some chemical reactions which are, ideally, the same for this VOC r at this concentration c_r. So, if we repeat N times the same experiment in the same conditions, the signature should be exactly the same and we should have N identical vectors. However, in the real world, there is always noise and the noise is random by nature. Therefore, the signature $\tilde{\boldsymbol{x}}_r$ from this noisy measurement is also random and after N experiments, we do not have N identical vectors but N vectors which just look alike. They look alike because they have been drawn from the same data distribution, the distribution of the VOC r at the concentration c_r. Let us assume that we measure this VOC in the same conditions but some days or weeks have elapsed since the first set of measurements. As we have seen in the previous sections, a drift may appear between the two data sets, meaning that the new signatures now differ from the older ones. This means that the data distribution has changed. If we could know the transformation between the new distribution and the old distribution, we could correct the new signatures and reuse all what we have previously learnt from the old distribution (*e.g.* the decision boundaries of a classifier). In the following, we use this notion by assuming that we know the general form of this transformation. To describe the data distributions, we use probability densities (a measure giving the probability that the signature lies in a given region of space).

Let $g(\tilde{\boldsymbol{x}}|r, t, c_r(t))$ be the probability density of the signature $\tilde{\boldsymbol{x}} \in \mathbb{R}^P$ of the VOC r, measured at time t, at concentration $c_r(t)$. Since we are interested in VOC classification, we assume that $\tilde{\boldsymbol{x}}$ has been normalized and we note $\boldsymbol{x}$ the concentration-free version of $\tilde{\boldsymbol{x}}$. We note

$f(\boldsymbol{x}|r,t)$ the probability density of the normalized signature $\boldsymbol{x}$, and we define as class r all the normalized signatures of the VOC r.

Under Fundamental Hypothesis, the drift of all VOCs is assumed to be related. Here, we assume that the distribution of each VOC r is simply shifted by a common vector $\boldsymbol{d}(t) \in \mathbb{R}^P$ weighted by a coefficient $\alpha_r(t) \in \mathbb{R}$ (depending on the VOC). Mathematically, it means that there exists a density $f_0(.|r)$ such that $f(\boldsymbol{x}|r,t) = f_0(\boldsymbol{x} + \alpha_r(t)\boldsymbol{d}(t)|r)$. We assume that $\boldsymbol{d}(t=0) = 0$, which means that the training set (Session 1) defines the starting position.

We can define the following drift model for a point $\boldsymbol{x}(t)$ at time t:

$$\boldsymbol{x}(t) \sim f_0(\boldsymbol{x} + \alpha_r(t)\boldsymbol{d}(t)|r) \tag{4.1}$$

Finally, we assume that the drift is a slow-varying process, which is a commonly used assumption ([Pad+10; DC+11]). This means that for a short time window Δ_t, we have $\boldsymbol{d}(t+\Delta_t) \approx \boldsymbol{d}(t)$ and $\alpha(t+\Delta_t) \approx \alpha(t)$. So, if Session 2 starts at $t = T_2$ and lasts only for few hours, we assume that each distribution r has been shifted by $\alpha_r(T_2)\boldsymbol{d}(T_2)$. In other words, noting $\boldsymbol{x}_1$ a sample from Session 1 (training session) and $\boldsymbol{x}_2$ a sample from Session 2 (testing session):

$$\boldsymbol{x}_2 \sim f_0(\boldsymbol{x}_1 + \alpha_r(T_2)\boldsymbol{d}(T_2)|r)$$

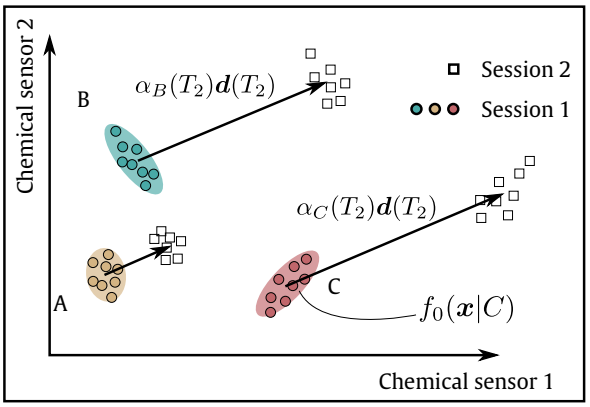


Fig. 4.5.: The drift model that we assume. Each class is translated along a common direction weighted by a coefficient depending on the class.

A simplified representation of the drift model is reported in Figure 4.5.

4.3.1 Simplified case: a frustratingly easy correction method

The drift model can be simplified by assuming that $\alpha_1(t) = ... = \alpha_R(t)$. This assumption states that the drift is a simple translation affecting the R distributions between Session 1 and samples $\in \mathcal{I}(\tau) = [\tau - \frac{\Delta_t}{2}, \tau + \frac{\Delta_t}{2}]$. In this case, a simple correction method can be derived by computing the expectations:

$$\begin{cases} \mathbb{E}(\boldsymbol{x}|t=0) & = \sum_{r=1}^{R} \mathbb{E}(\boldsymbol{x}|t=0,r)\pi_r(t=0) = \sum_{r=1}^{R} \boldsymbol{\mu}_r \pi_r(t=0) \\ \mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau)) & = \sum_{r=1}^{R} \mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau),r)\pi_r(t \in \mathcal{I}(\tau)) = \sum_{r=1}^{R} (\boldsymbol{\mu}_r + \boldsymbol{d}(\tau))\pi_r(t \in \mathcal{I}(\tau)) \end{cases}$$

with $\pi_r(t) = p(r|t)$ and μ_r the class centroid. So,

$$\mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau)) - \mathbb{E}(\boldsymbol{x}|t=0) = \boldsymbol{d}(\tau) + \sum_{r=1}^{R} \boldsymbol{\mu}_r(\pi_r(t \in \mathcal{I}(\tau)) - \pi_r(t=0))$$

Algorithm 7 Moving Recentering

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, \boldsymbol{X}(\tau) \in \mathbb{R}^{N(\tau) \times P}$
$\boldsymbol{\mu}_1 = \frac{1}{N_1} \boldsymbol{X}_1^{\mathrm{T}} \mathbf{1}_{N_1}, \quad \boldsymbol{\mu}(\tau) = \frac{1}{N(\tau)} \boldsymbol{X}(\tau)^{\mathrm{T}} \mathbf{1}_{N(\tau)}$
$\boldsymbol{d}(\tau) = \boldsymbol{\mu}(\tau) - \boldsymbol{\mu}_1$
$\boldsymbol{x}_{\text{cor}}(\tau) = \boldsymbol{x}(\tau) - \boldsymbol{d}(\tau)$

Then,

$$\mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau)) - \mathbb{E}(\boldsymbol{x}|t = 0) = \boldsymbol{d}(\tau) \iff \forall r,\ \pi_r(t \in \mathcal{I}(\tau)) = \pi_r(t = 0) \tag{4.2}$$

Equation (4.2) simply states that, if the frequency of occurrence, $\frac{N_r}{N}$, of each VOC r remains the same between Session 1 and the time window $\mathcal{I}(\tau)$, then the drift can be simply computed as the difference of the means. We call this scenario: **Equal Frequency of Occurrence Scenario** (EFOS)[3]. Equivalently, one can also simply center each data set individually. In this case, $\mathbb{E}(\boldsymbol{x}|t = 0) = 0$ and $\mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau)) = 0$ so $\boldsymbol{d}(\tau) = 0$.

We call this drift correction method: **Moving Recentering**. We note $\boldsymbol{X}(\tau) \in \mathbb{R}^{N(\tau) \times P}$ all the samples in the time window $[\tau - \frac{\Delta_t}{2}, \tau + \frac{\Delta_t}{2}]$, respecting the EFOS, and $\boldsymbol{x}(\tau)$ the sample at time τ. The algorithm for correcting $\boldsymbol{x}(\tau)$ is reported in Algorithm 7.

If Session 2 lasts for few hours, then $\boldsymbol{X}(\tau) \in \mathbb{R}^{N(\tau) \times P}$ may correspond to the entire session (all samples will be in the time window $[\tau - \frac{\Delta_t}{2}, \tau + \frac{\Delta_t}{2}]$). In this case, in Algorithm 7, we can replace $\boldsymbol{X}(\tau)$ by the data coming from Session 2, $\boldsymbol{X}_2 \in \mathbb{R}^{N_2 \times P}$. Algorithm 7 is then just a recentering of data from Session 2 on data from Session 1.

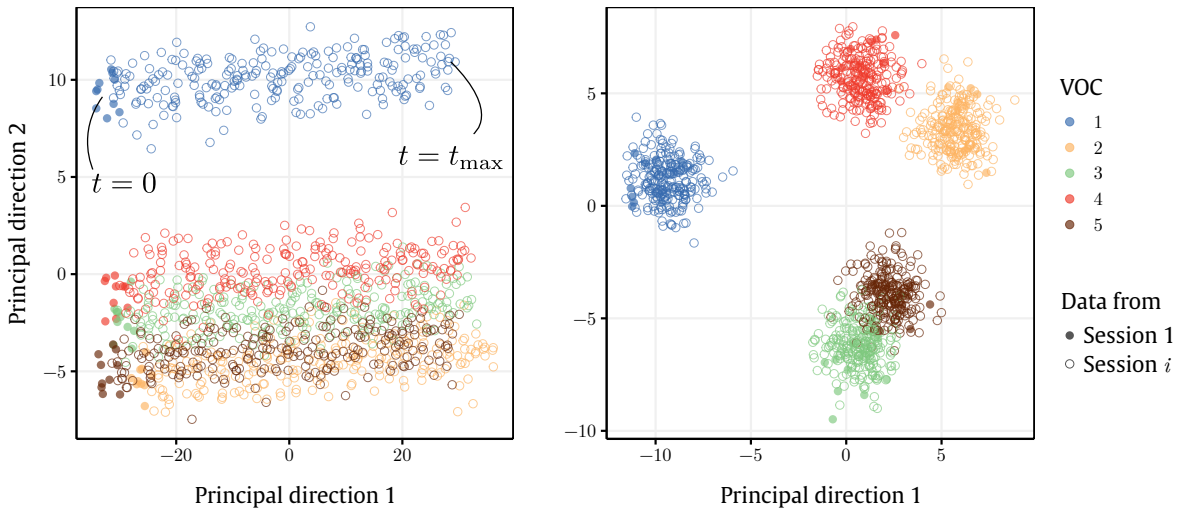
Algorithm 7 is somewhat trivial but is clearly neglected in the literature. To illustrate this remark, let us look closely at the paper of Di Carlo et al. [DC+11] from which we have previously detailed the correction method (see Section 4.2.3, Algorithm 6).

In their article, Di Carlo et al. tested their correction method with two datasets, one artificial dataset and one real dataset. The artificial dataset can be easily replicated:

$$\begin{cases} \forall r \in [\![1, R]\!],\ \boldsymbol{\mu}_r \sim \mathcal{N}(\mathbf{0}_P, \frac{\beta^2}{2P} \boldsymbol{I}_P) & \text{with } \beta = 12, R = 5, P = 6 \\ \tilde{\boldsymbol{d}} \sim \mathcal{U}(0, 1) \text{ and } \boldsymbol{d} = \frac{\tilde{\boldsymbol{d}}}{\sqrt{\tilde{\boldsymbol{d}}^{\mathrm{T}} \tilde{\boldsymbol{d}}}} & \\ \forall t \in [\![0, t_{\max}]\!],\ \boldsymbol{x}_r(t) \sim \mathcal{N}(\boldsymbol{\mu}_r + \frac{t}{h} \boldsymbol{d}, \boldsymbol{I}_P) & \text{with } t_{\max} = 250, h = 4 \end{cases} \tag{4.3}$$

The simulation settings of Di Carlo et al. [DC+11] clearly fulfill the requirements for applying Algorithm 7: drift is a simple translation independent of the class and we are in the Equal Frequency of Occurence Scenario (EFOS). Only the parameter h has been decreased (by a factor 10) compared to the study of Di Carlo et al., in order to include more drift. In Figure 4.6, we represent the projection of the artificial dataset onto the 2 first Principal directions,

[3] A way to relax the assumption of Equal Frequency of Occurrence Scenario can be to consider the Calibrant Scenario. If VOC 1 is the calibrant, then $\mathbb{E}(\boldsymbol{x}|t \in \mathcal{I}(\tau), r = 1) - \mathbb{E}(\boldsymbol{x}|t = 0, r = 1) = \boldsymbol{d}(\tau)$. So the drift direction is simply estimated as the difference between the mean of the calibrating samples from Session 1 and the mean of the calibrating samples from the time window $\mathcal{I}(\tau)$.



(a) Before correction

(b) After Moving Recentering

Fig. 4.6.: Artificial dataset used by Di Carlo et al. [DC+11], generated according to the simulation settings in Eq. (4.3). The two first Principal components of the data (a) before correction and (b) after applying Moving Recentering (Algorithm 7). $\boldsymbol{X}_1$ corresponds to the solid points (10 first points of each VOC).

for a random realization. Figure 4.6a corresponds to the raw data (without any correction) and Figure 4.6b corresponds to the data after applying Algorithm 7. Our algorithm has been applied by considering only the 10 first samples of each VOC (for $\boldsymbol{X}_1$) and we set $\Delta_t = 5$. The results are clear and the drift has been completely corrected.

The real dataset used in [DC+11] is not publicly available so we simply extracted the figure representing the two first Principal components of the real dataset from the article. It corresponds to the measurements of 5 different VOCs over 1 month by using 6 MOS sensors. The drift in Figure 4.7 matches the simulated one, as pointed out by the authors themselves (see [DC+11]). So it is likely that we can replicate the results obtained in simulation with our method. It is noteworthy that Algorithm 7 is simpler than Algorithm 6 by far (the correction method proposed by Di Carlo et al.) and requires no classification step. Besides, the last point indicates that our method will adapt in case of a Discontinuous drift, contrary to the algorithm of Di Carlo et al.

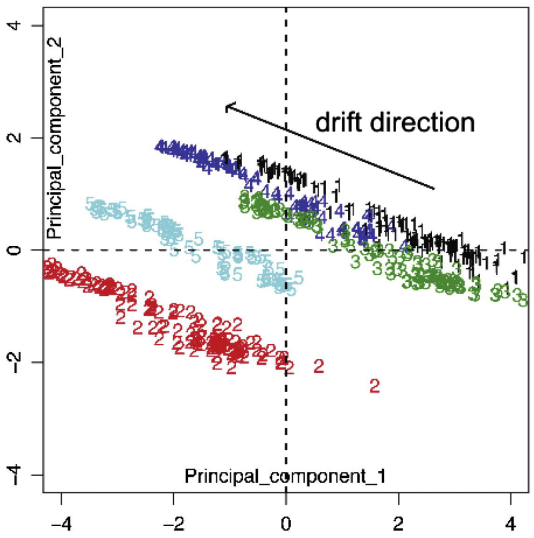
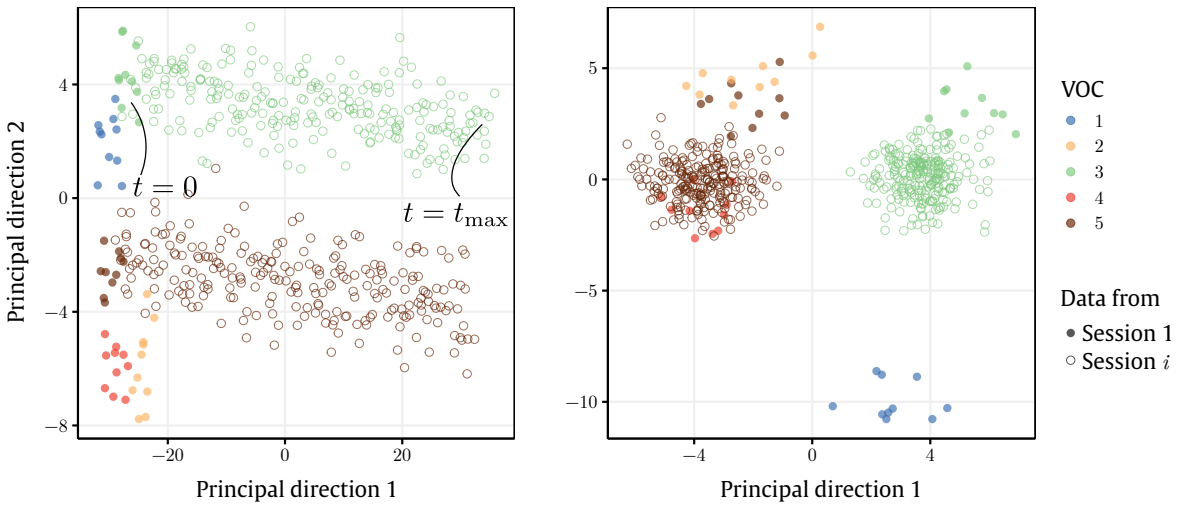


Fig. 4.7.: The two first principal components of the real dataset of Di Carlo et al. Figure retaken from [DC+11].



(a) Before correction

(b) After Moving Recentering

Fig. 4.8.: Artificial dataset used by Di Carlo et al. [DC+11] when we are no longer in the EFOS. The two first principal components of the data (a) before correction and (b) after Moving Recentering (Algorithm 7). $\boldsymbol{X}_1$ corresponds to the solid points (10 first points of each class).

4.3.2 Gaussian case: Expectation-Maximization Component Correction (EMC^2)

In practice, a training set with R VOCs will be generated during Session 1 with a certain frequency of occurrence $\pi_r(t = 0)$ but it is unlikely that these π_r will remain the same during next sessions. In the best case, the R VOCs will reappear in different proportions but it is more realistic to expect only $\gamma < R$ VOCs. In other words, the Equal Frequency of Occurrence Scenario is somewhat unrealistic. To illustrate this fundamental limitation of Algorithm 7, we can use the simulation settings in Eq. (4.3) and remove 3 classes (so $\gamma = 2$) for the sessions coming after Session 1. In Figure 4.8, we set: $\pi_1(t \neq 0) = \pi_2(t \neq 0) = \pi_4(t \neq 0) = 0$ and $\pi_3(t \neq 0) = \pi_5(t \neq 0) = 0.5$. The consequences are clear and the correction no longer works. In fact, Figure 4.8 shows that there exists a latent problem of classification which makes the task complicated. Indeed, if you know that only classes 3 and 5 are appearing in the next sessions, then we could apply Algorithm 7 restricting the training set to classes 3 and 5 and the problem would be solved. Unfortunately, this is not the case, so we have to manage the classification problem and the drift estimation at the same time.

In the following, we define an Expectation-Maximization algorithm for dealing with the drift model in Equation (4.1). For a set of observations $\boldsymbol{X}(\tau)$ acquired in a time window $\mathcal{I}(\tau) = [\tau - \frac{\Delta_t}{2}, \tau + \frac{\Delta_t}{2}]$, the algorithm aims at estimating both the drift direction $\boldsymbol{d}(\tau)$, the class weights $\alpha_r(\tau)$, the frequencies of occurrence $\pi_r(\tau)$ and the labels $\boldsymbol{\ell}(\tau)$.

Expectation-Maximization (EM) algorithm is an iterative method for finding maximum likelihood solutions of statistical models depending on latent variables [DLR77; Bis06]. Latent variables are some variables which cannot be directly observed. In our special case, the latent variables are the labels $\boldsymbol{\ell}(\tau)$ of $\boldsymbol{X}(\tau)$ and the parameters are the drift direction

$\boldsymbol{d}(\tau)$ and the weights $\alpha_r(\tau)$. The main difficulty is that we cannot estimate $\boldsymbol{\ell}(\tau)$ without the knowledge of $\boldsymbol{d}(\tau)$ and $\alpha_r(\tau)$, and we cannot estimate $\boldsymbol{d}(\tau)$ and $\alpha_r(\tau)$ without the knowledge of $\boldsymbol{\ell}(\tau)$. To solve this issue, an EM iteration proceeds in two steps: the Expectation step and the Maximization step. The Expectation step will estimate the latent variables by considering that the parameters are known and the Maximization step will do the reverse. These 2 steps are then alternated until convergence.

To describe the EM formulation, we assume that each VOC r is drawn from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{S}_r)$, with $\boldsymbol{\mu}_r \in \mathbb{R}^P$ the class centroid and $\boldsymbol{S}_r \in \mathbb{R}^{P\times P}$ the class covariance matrix. $\boldsymbol{\mu}_r$ and $\boldsymbol{S}_r$ can be estimated from the training set $\boldsymbol{X}_1 \in \mathbb{R}^{N_1\times P}$ (Session 1), by noting $\boldsymbol{X}_{r,1} \in \mathbb{R}^{N_{r,1}\times P}$ the submatrix containing the VOC r:

$$\hat{\boldsymbol{\mu}}_r = \frac{1}{N_{r,1}} \boldsymbol{X}_{r,1}^{\mathrm{T}} \mathbf{1}_{N_{r,1}} \quad \text{and} \quad \hat{\boldsymbol{S}}_r = \frac{1}{N_{r,1}-1} (\boldsymbol{X}_{r,1} - \mathbf{1}_{N_{r,1}} \hat{\boldsymbol{\mu}}_r^{\mathrm{T}})^{\mathrm{T}} (\boldsymbol{X}_{r,1} - \mathbf{1}_{N_{r,1}} \hat{\boldsymbol{\mu}}_r^{\mathrm{T}})$$

In the following, to lighten the notation, $\boldsymbol{\mu}_r$ and $\boldsymbol{S}_r$ are simply replaced by their estimates $\hat{\boldsymbol{\mu}}_r$ and $\hat{\boldsymbol{S}}_r$.

For a sample $\boldsymbol{x}(\tau)$ acquired with other samples in a time window $\mathcal{I}(\tau) = [\tau - \frac{\Delta_t}{2}, \tau + \frac{\Delta_t}{2}]$, the drift model states that:

$$p(\boldsymbol{x}(\tau)|\ell(\tau) = r) = \mathcal{N}(\boldsymbol{\mu}_r + \alpha_r(\tau)\boldsymbol{d}(\tau), \boldsymbol{S}_r) \tag{4.4}$$

To lighten the notation, we remove the time τ from the notations. So, we have a set of observations $\boldsymbol{X}(\tau) \in \mathbb{R}^{N(\tau)\times P}$ which is abbreviated $\boldsymbol{X} \in \mathbb{R}^{N\times P}$. One signature in $\boldsymbol{X}$ is noted $\boldsymbol{x}_n \in \mathbb{R}^P$. The weights α_r and the class probabilities π_r are stacked into vectors $\boldsymbol{\alpha} \in \mathbb{R}^R$ and $\boldsymbol{\pi} \in \mathbb{R}^R$.

Expectation step. In this step, we assume that $\boldsymbol{d}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ are known. Then, we can estimate the *a posteriori* probability $p(\ell_n = r|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ that $\boldsymbol{x}_n$ belongs to the class r:

$$\begin{aligned} p(\ell_n = r|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) &\overset{\text{Bayes rule}}{=} \frac{p(\ell_n = r)p(\boldsymbol{x}_n|\ell_n = r, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{p(\boldsymbol{x}_n|\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})} \\ &\overset{\text{Marginalization}}{=} \frac{p(\ell_n = r)p(\boldsymbol{x}_n|\ell_n = r, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{\sum_{k=1}^{R} p(\ell_n = k)p(\boldsymbol{x}_n|\ell_n = k, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})} \end{aligned}$$

$$\boxed{p(\ell_n = r|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) = \frac{\pi_r \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_r + \alpha_r \boldsymbol{d}, \boldsymbol{S}_r)}{\sum_{k=1}^{R} \pi_k \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_k + \alpha_k \boldsymbol{d}, \boldsymbol{S}_k)}} \tag{4.5}$$

For a given r, the closer to 1 $p(\ell_n = r|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ is, the more likely it is that $\boldsymbol{x}_n$ belongs to class r.

Maximization step. In this step, we assume that $p(\ell_n = r|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ is known. Then, we can estimate each parameter by maximizing the likelihood. The likelihood is the probability to have the set of observations $\boldsymbol{X}$ given the parameters $\boldsymbol{d}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. It is defined as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X}) &= p(\boldsymbol{X}|\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \\ &= p(\boldsymbol{x}_1, ..., \boldsymbol{x}_N|\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \\ &\overset{\text{independent and identically distributed}}{=} \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})\end{aligned}$$

We usually prefer to work with the log-likelihood $\mathcal{H}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X}) = \ln(\mathcal{L}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X}))$:

$$\begin{aligned}\mathcal{H}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X}) &= \sum_{n=1}^{N} \ln(p(\boldsymbol{x}_n|\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})) \\ &= \sum_{n=1}^{N} \ln(\sum_{k=1}^{R} \pi_k \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_k + \alpha_k \boldsymbol{d}, \boldsymbol{S}_k))\end{aligned}$$

Then, to find the parameters which have likely generated the set of observations $\boldsymbol{X}$, we can maximize $\mathcal{H}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X})$ with respect to these parameters. To do this we simply write the derivatives with respect to each parameter and set them to zero.

We note $\mathcal{N}(k) = \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_k + \alpha_k \boldsymbol{d}, \boldsymbol{S}_k)$ and $\rho_{nk} = p(\ell_n = k|\boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})$.

① **Estimation of $\boldsymbol{d}$.**

$$\frac{\partial \mathcal{H}}{\partial \boldsymbol{d}} = \sum_{n=1}^{N} \sum_{k=1}^{R} \underbrace{\frac{\pi_k \mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)}}_{\rho_{nk}} \underbrace{\frac{\partial}{\partial \boldsymbol{d}}(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \alpha_k \boldsymbol{d})^{\mathrm{T}} \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \alpha_k \boldsymbol{d}))}_{\alpha_k \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \alpha_k^2 \boldsymbol{S}_k^{-1} \boldsymbol{d}}$$

So,

$$\frac{\partial \mathcal{H}}{\partial \boldsymbol{d}} = 0 \iff \sum_{n,k} \rho_{nk} \alpha_k^2 \boldsymbol{S}_k^{-1} \boldsymbol{d} = \sum_{n,k} \rho_{nk} \alpha_k \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

$$\boxed{\hat{\boldsymbol{d}} = \Lambda^{-1} \sum_{n,k} \rho_{nk} \alpha_k \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)} \tag{4.6}$$

with $\boxed{\Lambda = \sum_k \alpha_k^2 \boldsymbol{S}_k^{-1} \sum_n \rho_{nk}}$

② **Estimation of $\boldsymbol{\alpha}$.**

$$\frac{\partial \mathcal{H}}{\partial \alpha_k} = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)}}_{\rho_{nk}} \underbrace{\frac{\partial}{\partial \alpha_k}(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \alpha_k \boldsymbol{d})^{\mathrm{T}} \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k - \alpha_k \boldsymbol{d}))}_{\boldsymbol{d}^{\mathrm{T}} \boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \alpha_k \boldsymbol{d}^{\mathrm{T}} \boldsymbol{S}^{-1} \boldsymbol{d}}$$

So,

$$\frac{\partial \mathcal{H}}{\partial \alpha_k} = 0 \iff \sum_n \rho_{nk}\alpha_k \boldsymbol{d}^{\mathrm{T}}\boldsymbol{S}_k^{-1}\boldsymbol{d} = \sum_n \rho_{nk}\boldsymbol{d}^{\mathrm{T}}\boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

If $\sum_n \rho_{nk} \neq 0$ then,

$$\boxed{\forall k \in [\![1, R]\!], \quad \hat{\alpha}_k = \frac{\sum_n \rho_{nk}\boldsymbol{d}^{\mathrm{T}}\boldsymbol{S}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{S}_k^{-1}\boldsymbol{d}\sum_n \rho_{nk}}} \tag{4.7}$$

Otherwise $\hat{\alpha}_k = 0$.

③ **Estimation of $\boldsymbol{\pi}$.** Since vector $\boldsymbol{\pi}$ contains probabilities, $\sum_{k=1}^{R} \pi_k = 1$. To incorporate this constraint in the optimization, we build the Lagrangian $\Gamma(\boldsymbol{\pi}, \lambda)$:

$$\Gamma(\boldsymbol{\pi}, \lambda) = \mathcal{H}(\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\boldsymbol{X}) + \lambda(\sum_{k=1}^{R} \pi_k - 1)$$

For maximizing $\Gamma(\boldsymbol{\pi}, \lambda)$, we compute the derivative and set it to zero,

$$\frac{\partial \Gamma}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} + \lambda \tag{4.8}$$

Hence,

$$\begin{aligned}
\frac{\partial \Gamma}{\partial \pi_k} = 0 &\iff \sum_{n=1}^{N} \frac{\mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} + \lambda = 0 \\
&\overset{\times \pi_k}{\iff} \sum_{n=1}^{N} \pi_k \frac{\mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} + \pi_k \lambda = 0 \\
&\overset{\text{Sum over k}}{\iff} \sum_{n=1}^{N} \sum_{k=1}^{R} \pi_k \frac{\mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} + \sum_{k=1}^{R} \pi_k \lambda = 0 \\
&\iff \sum_{n=1}^{N} \frac{\sum_{k=1}^{R} \pi_k \mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} + \lambda = 0 \\
&\iff \lambda = -N
\end{aligned}$$

To conclude, inject $\lambda = -N$ in Equation (4.8),

$$\sum_{n=1}^{N} \frac{\mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} - N = 0 \overset{\times \pi_k}{\iff} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(k)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(i)} - \pi_k N = 0$$

$$\boxed{\forall k \in [\![1, R]\!], \quad \hat{\pi}_k = \frac{\sum_n \rho_{nk}}{N}} \tag{4.9}$$

Equations (4.5) and (4.9) are quite classical in EM algorithms for mixture estimation. However, Equations (4.6) and (4.7) are derived from the drift model that we used. There is an obvious indetermination in our method: a scale ambiguity between $\boldsymbol{\alpha}$ and $\boldsymbol{d}$. Indeed, for any $\lambda \in \mathbb{R}^*$: $\forall r \in [\![1, R]\!]$, $\frac{\alpha_r}{\lambda} \times \lambda \boldsymbol{d} = \alpha_r \boldsymbol{d}$. To remove this ambiguity, we simply normalize $\boldsymbol{d}$ after Equation (4.6).

After convergence, we need a label ℓ_n to correct the sample $\boldsymbol{x}_n$ (the weight depends on the class). We take the class which maximizes the *a posteriori* probability:

$$\ell_n = \arg\max_r p(\ell_n = r | \boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \tag{4.10}$$

Once we have estimated the label ℓ_n of the sample $\boldsymbol{x}_n$, we can correct it:

$$\boldsymbol{x}_{n,\text{cor}} = \boldsymbol{x}_n - \alpha_{\ell_n} \boldsymbol{d} \tag{4.11}$$

It is interesting to note that our EM-based algorithm removes one component from $\boldsymbol{x}_n$ which is $\alpha_{\ell_n} \boldsymbol{d}$. It is reminiscent of some methods from the literature called Component Correction (see Section 4.2). So we call this method: **Expectation-Maximization Component Correction**, abbreviated **EMC**^{2}. The algorithm is reported in Algorithm 8.

For an easier implementation, we recall the main formulas, assuming that we are at iteration i:

- Equation (4.5) $\rho_{nr}^{(i)} = p^{(i)}(\ell_n = r | \boldsymbol{x}_n, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}) = \dfrac{\pi_k^{(i-1)} \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_r + \alpha_r^{(i-1)} \boldsymbol{d}^{(i-1)}, \boldsymbol{S}_r)}{\sum_{k=1}^{R} \pi_k^{(i-1)} \mathcal{N}(\boldsymbol{x}_n, \boldsymbol{\mu}_k + \alpha_k^{(i-1)} \boldsymbol{d}^{(i-1)}, \boldsymbol{S}_k)}$
- Equation (4.6) $\boldsymbol{d}^{(i)} = \Lambda^{-1} \sum_{n,k} \rho_{nk}^{(i)} \alpha_k^{(i-1)} \boldsymbol{S}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$
  with $\Lambda = \sum_k (\alpha_k^{(i-1)})^2 \boldsymbol{S}_k^{-1} \sum_n \rho_{nk}^{(i)}$
- Equation (4.7) $\forall k \in [\![1, R]\!]$, $\alpha_k^{(i)} = \dfrac{\sum_n \rho_{nk}^{(i)} \boldsymbol{d}^{(i)^{\mathrm{T}}} \boldsymbol{S}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\boldsymbol{d}^{(i)^{\mathrm{T}}} \boldsymbol{S}_k^{-1} \boldsymbol{d}^{(i)} \sum_n \rho_{nk}^{(i)}}$
- Equation (4.9) $\forall k \in [\![1, R]\!]$, $\pi_k^{(i)} = \dfrac{\sum_n \rho_{nk}^{(i)}}{N}$

Some possible variants of Algorithm 8 can be implemented effortlessly:

- As an extension of Algorithm 7. In Algorithm 7, we assume that $\boldsymbol{\alpha} = \mathbf{1}_R$ but Algorithm 7 will likely fail if we are not in the EFOS. Algorithm 8 does not require such a hypothesis and the estimation of $\boldsymbol{\alpha}$ could be simply skipped (setting $\boldsymbol{\alpha} = \mathbf{1}_R$).
- To deal with the case where Session 1 contains few samples compared to the number of chemical sensors. In this case, the R covariance matrices $\boldsymbol{S}_r$ could be underestimated. In the worst case, this could lead to numerical instabilities since the matrices $\boldsymbol{S}_r$ are often inversed in our algorithm. To prevent this issue, two solutions can be considered:

Algorithm 8 Expectation-Maximization Component Correction (EMC^2)

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, \boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}, \boldsymbol{X} \in \mathbb{R}^{N \times P}$
Initialization: $\boldsymbol{d}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\pi}^{(0)}$
Compute the R centroids $\boldsymbol{\mu}_r$ and the R covariance matrices $\boldsymbol{S}_r$ from $\boldsymbol{X}_1$
while $\boldsymbol{d}^{(i)}$ unchanged **do**
  # Expectation step
  Compute $p^{(i)}(\ell_n = r|\boldsymbol{x}_n)$ from Equation (4.5), $\forall n \in [\![1, N]\!]$, $\forall r \in [\![1, R]\!]$
  # Maximization step
  Compute $\boldsymbol{\pi}^{(i)}$ from Equation (4.9)
  Compute $\boldsymbol{d}^{(i)}$ from Equation (4.6)
  $\boldsymbol{d}^{(i)} = \frac{\boldsymbol{d}^{(i)}}{\sqrt{\boldsymbol{d}^{(i)\mathrm{T}}\boldsymbol{d}^{(i)}}}$
  if $\pi_k^{(i)} \geqslant \frac{1}{N}$ **then**
    Compute $\alpha_k^{(i)}$ from Equation (4.7)
  else
    $\alpha_k^{(i)} = 0$
  end if
end while
 # Drift correction
$\forall n \in [\![1, N]\!]$, $\ell_n = \arg\max_r p(\ell_n = r|\boldsymbol{x}_n)$
$\forall n \in [\![1, N]\!]$, $\boldsymbol{x}_{n,\text{cor}} = \boldsymbol{x}_n - \alpha_{\ell_n}\boldsymbol{d}$

- Approximate $\boldsymbol{S}_r$ as a diagonal matrix $\boldsymbol{\Sigma}_r$.
- Set the same covariance matrix for all classes: $\boldsymbol{S} = \sum_{r=1}^{R} \boldsymbol{S}_r$.

4.3.3 Initialization of EMC^2

EMC^2 is quite sensitive to the initial set of parameters $\boldsymbol{d}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\pi}^{(0)}$. As a trivial initialization, we can take the parameters derived from Algorithm 7:

$$\begin{cases} \boldsymbol{d}^{(0)} &= \frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\mathbf{1}_N - \frac{1}{N_1}\boldsymbol{X}_1^{\mathrm{T}}\mathbf{1}_{N_1} \\ \boldsymbol{\alpha}^{(0)} &= \mathbf{1}_R \\ \boldsymbol{\pi}^{(0)} &= \frac{1}{R}\mathbf{1}_R \end{cases} \tag{4.12}$$

However, this initialization will likely lead to a bad correction if the initial parameters are far from the real ones. Instead, we propose a heuristic based on orthogonalizations and k-NN (k-Nearest Neighbors) to find better initial parameters.

The main idea is that a simple orthogonalization against $\boldsymbol{d}$ will align $\boldsymbol{X}_1$ and $\boldsymbol{X}$ and information about $\boldsymbol{d}$ must be contained in some pairwise directions between some samples in $\boldsymbol{X}_1$ (here, we use the centroids $\boldsymbol{\mu}_r$ since we know the labels $\boldsymbol{\ell}_1$) and some samples in $\boldsymbol{X}$. We start by selecting some random directions going from each centroid $\boldsymbol{\mu}_r$ of $\boldsymbol{X}_1$ to some random samples of $\boldsymbol{X}$. For each random direction, we project $\boldsymbol{X}$ and $\boldsymbol{X}_1$ in the subspace orthogonal to the candidate direction. These versions of $\boldsymbol{X}$ and $\boldsymbol{X}_1$ are called $\boldsymbol{X}_\perp$ and $\boldsymbol{X}_{1,\perp}$. A good direction would then align $\boldsymbol{X}_\perp$ and $\boldsymbol{X}_{1,\perp}$, meaning that we would have only R different clusters. A bad direction would however create additional clusters in the subspace. To check if

$\boldsymbol{X}_\perp$ and $\boldsymbol{X}_{1,\perp}$ are aligned or not, a good figure of merit is the within-class covariance matrix that we already used in Chapter 3 (Section 3.2.3). Note $\check{\boldsymbol{X}}_\perp$ the combination of the two datasets $\boldsymbol{X}_\perp$ and $\boldsymbol{X}_{1,\perp}$:

$$\boldsymbol{S}_w = \sum_{r=1}^{R} \sum_{n \in \text{VOC } r} (\check{\boldsymbol{x}}_{n,\perp} - \boldsymbol{\mu}_{r,\perp})(\check{\boldsymbol{x}}_{n,\perp} - \boldsymbol{\mu}_{r,\perp})^{\mathrm{T}}$$

with $\boldsymbol{\mu}_{r,\perp}$ the centroid of the class r in the orthogonal subspace.

Algorithm 9 Heuristic for initializing EMC^2

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}, \boldsymbol{\ell}_1 \in \mathbb{R}^{N_1}, \boldsymbol{X} \in \mathbb{R}^{N \times P}, A \in \mathbb{N}, k_{\text{kNN}} \in \mathbb{N}$
Compute the centroids $\boldsymbol{\mu}_r(\boldsymbol{X}_1)$ from $\boldsymbol{X}_1$
Select randomly a set of A samples $\boldsymbol{X}_A \in \mathbb{R}^{A \times P}$ in $\boldsymbol{X}$
Compute all the $B = A \times R$ possible directions with unit-norm between each sample in $\boldsymbol{X}_A$ and each centroid $\boldsymbol{\mu}_r(\boldsymbol{X}_1)$
for i in $1 : B$ **do**
  $\boldsymbol{X}_{1,\perp} = \boldsymbol{X}_1(\boldsymbol{I}_P - \boldsymbol{d}_i\boldsymbol{d}_i^{\mathrm{T}})$
  $\boldsymbol{X}_\perp = \boldsymbol{X}(\boldsymbol{I}_P - \boldsymbol{d}_i\boldsymbol{d}_i^{\mathrm{T}})$
  $\boldsymbol{\ell}_i = k\text{NN}(\boldsymbol{X}_{1,\perp}, \boldsymbol{\ell}_1, \boldsymbol{X}_\perp, k_{\text{kNN}})$ *# Classify in the subspace orthogonal to the candidate direction*
  Merge $\boldsymbol{X}_{1,\perp}$ and $\boldsymbol{X}_\perp$ as $\check{\boldsymbol{X}}_\perp$, $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_i$ as $\check{\boldsymbol{\ell}}$
  Compute the score w_i with $(\check{\boldsymbol{X}}_\perp, \check{\boldsymbol{\ell}})$ *# w_i is the trace of the within-class covariance matrix*
end for
$I = \arg\min_i w_i$
$\boldsymbol{\ell} = \boldsymbol{\ell}_I, \quad \boldsymbol{d}^{(0)} = \boldsymbol{d}_I, \quad \alpha_r^{(0)} = (\boldsymbol{\mu}_r(\boldsymbol{X}) - \boldsymbol{\mu}_r(\boldsymbol{X}_1))^{\mathrm{T}}\boldsymbol{d}^{(0)}, \quad \pi_r^{(0)} = \frac{\#\{\boldsymbol{\ell}_I = r\}}{N}$

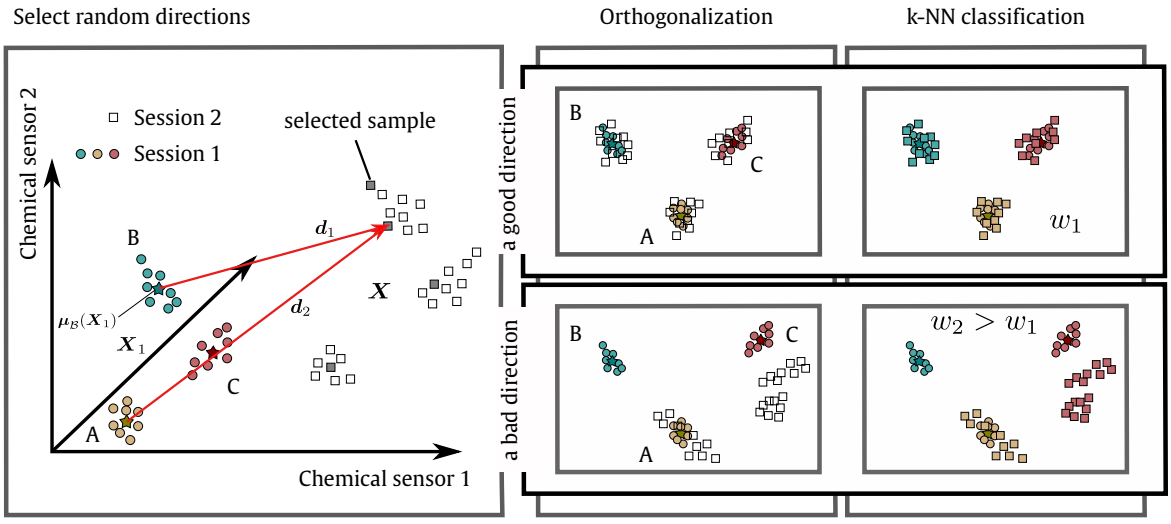


Fig. 4.9.: Graphical representation of the initialization procedure for EMC^2 ($A = 4, R = 3$). We first select random directions from the centroids of $\boldsymbol{X}_1$ to some random samples of $\boldsymbol{X}$. Then, for each candidate direction, we orthogonalize $\boldsymbol{X}_1$ and $\boldsymbol{X}$ against the direction. In the subspace, we classify $\boldsymbol{X}_\perp$ with a k-NN ($k = k_{\text{kNN}}$) trained on $\boldsymbol{X}_{1,\perp}$. Finally we compute a score w as the trace of the within-class covariance matrix in the subspace taking both sessions. Bad directions is assumed to have a greater value w. In the figure, $\boldsymbol{d}_1$ is clearly a better initial direction than $\boldsymbol{d}_2$.

$\boldsymbol{S}_w$ evaluates how far from each other are the samples belonging to a same class. The greater the within-class covariance is, the more spread-out the classes are. To have only a single number, we take $w = \text{Tr}(\boldsymbol{S}_w)$. The lower w is, the more $\boldsymbol{X}_\perp$ and $\boldsymbol{X}_{1,\perp}$ are aligned. However, the within-class covariance matrix needs labels and we do not have any label for $\boldsymbol{X}_\perp$, so a classification step is required. For that, we simply apply a k-NN in the orthogonal subspace of $\boldsymbol{d}$ to estimate the labels $\boldsymbol{\ell}$ of $\boldsymbol{X}_\perp$ based on $(\boldsymbol{X}_{1,\perp}, \boldsymbol{\ell}_1)$. The initialization procedure is described in Algorithm 9 and a graphical representation is reported in Figure 4.9.

4.4 Simulation

To validate EMC^2 (Algorithm 8), we start with an artificial data set. In this artificial data set, a training set (Session 1) is generated with R VOCs and test sets (future sessions) are generated according to the drift model in Equation (4.1) with $\gamma \leqslant R$ VOCs. These simulations aim at answering two questions:

- Can EMC^2 adapt to a variable number of VOCs γ?
- Can EMC^2 cope with Continuous and Discontinuous drift?

4.4.1 Simulation settings

For the generation of the artificial data set, we reuse and complicate the simulation settings of Di Carlo et al. [DC+11] which were introduced in Equation (4.3). Simulations start with the generation of a training set $\boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}$ with no drift and with R VOCs. Then, we generate data with drift according to the model in Equation (4.1) and only $\gamma \leqslant R$ VOCs are simulated (for the same training set, we test for γ from R to 2). Finally, a differentiation is done if the drift is Continuous, or rather Discontinuous.

Our method only implies that the drift is additive, slow-varying and along a direction $\boldsymbol{d}(t)$ independent of the VOC (only the weight factor is). So, $\boldsymbol{d}(t)$ is not constrained to be constant or linear over time, and can be rather complex. To introduce a non-linearity in the drift over time, we simulate the direction $\boldsymbol{d}(t)$ with an arbitrary non-linear function: $\boldsymbol{d}(t) = \frac{t}{h}\boldsymbol{d}_0 + \frac{t}{h}\sin(\frac{t}{h}\boldsymbol{d}_0)$, $\sin$ is applied component-by-component, $\boldsymbol{d}_0$ is a random direction and h is a hyperparameter. The simulation settings are reported in Table 4.2.

Since the simulations are random, the generation of the clusters could be not easily separable. In this case, we may not evaluate only the drift correction but also the initial discrimination between classes. To prevent this issue, we discard simulations for which any pairwise distance between centroids is lower than $\frac{\beta}{2}$ [DC+11]. The simulations are repeated 100 times.

4.4.2 Results

To evaluate the drift correction, we use a classification rate as criterion. In practice, one wants to recognise the R VOCs based only on the first session (here, Session 1 contains only 10 samples/class), no matter what the drift. So this criterion seems appropriate and realistic. EMC^2 (Algorithm 8) already includes a classification step for selecting the weight α_r, so we

Parameters for Session 1

$$\forall r \in [\![1, R]\!],\ \boldsymbol{\mu}_r \sim \mathcal{N}(\mathbf{0}_P, \tfrac{\beta^2}{2P}\boldsymbol{I}_P)$$

$$\forall r \in [\![1, R]\!],\ \boldsymbol{x}_r \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{I}_P),\ \boldsymbol{X}_1 \in \mathbb{R}^{N_1 \times P}$$

$$\beta = 12, R = 6, P = 4, N_1 = 60$$

Parameters for next Sessions

$$\tilde{\boldsymbol{d}}_0 \sim \mathcal{U}(0, 1) \text{ and } \boldsymbol{d}_0 = \frac{\tilde{\boldsymbol{d}}_0}{\sqrt{\tilde{\boldsymbol{d}}_0{}^{\top}\tilde{\boldsymbol{d}}_0}}$$

$$\boldsymbol{\alpha} \sim \mathcal{U}(0.5, 3)$$

$$\forall r \in [\![1, R]\!],\ \text{if } r \leqslant \gamma \text{ then } \pi_r = \tfrac{1}{\gamma},\ \text{otherwise } \pi_r = 0$$

$$\forall r \in [\![1, \gamma]\!],\ \forall t \in \mathcal{I},\ \boldsymbol{x}_r(t) \sim \mathcal{N}(\boldsymbol{\mu}_r + \alpha_r \tfrac{t}{h}(\boldsymbol{d}_0 + \sin(\tfrac{t}{h}\boldsymbol{d}_0)), \boldsymbol{I}_P)$$

$$t_{\max} = 251, h = 30, \Delta_t = 8, \gamma \text{ is variable}$$

Continuous drift	Discontinuous drift
$\mathcal{I} = [\![0, t_{\max}]\!]$	$\mathcal{I} = [\![t_{\max} - \frac{\Delta_t}{2}, t_{\max} + \frac{\Delta_t}{2}]\!]$

Tab. 4.2.: Simulation settings.

directly compare these labels to the ground truth. EMC^2 is initialized with the procedure described in Algorithm 9.

Performance of EMC^2 is also compared to three other methods. First, we test the *Raw* method which does not apply any correction to the new measurements. This method is straightforward: we learn a linear SVM from Session 1 and next samples are directly classified. The classification rates with Raw method will be used as base scores for the other methods. The second method is the method of Di Carlo et al. [DC+11] which is described in Algorithm 6[4]. This method also includes a classification step, so we directly use the generated labels. Finally, we also use PCA-CC from Artursson et al. [Art+00] (see Algorithm 2). Comparison with PCA-CC is a bit unfair since this method requires much more information, in particular calibrating samples. But the scores of PCA-CC can be seen as the "best" values. The methods considering the Multi-Session Scenario are not used here since the first sessions will not contain enough information about the future directions ($\boldsymbol{d}(t)$ is rather complex).

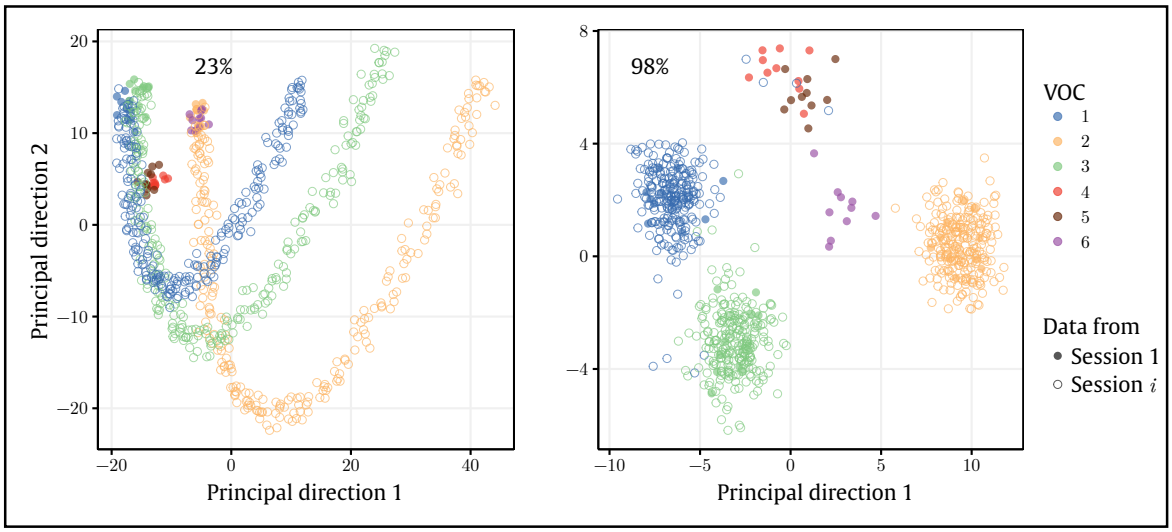
For PCA-CC, a number of components must be chosen. To automatically tune this parameter, we take the minimal number of components to explain at least 80% of the variance of $\boldsymbol{X}_{\text{cal}}$ (the matrix containing the calibrating samples) [Art+00]. For the classification, the corrected samples are classified with a linear SVM learnt from Session 1. VOC 1 is always chosen as the reference VOC.

Both EMC^2 and the method of Di Carlo et al. can be applied on time windows. So we segment $\boldsymbol{X}$ as a series of non-overlapping windows of size $\Delta_t + 1 = 9$. For the Continuous drift, it

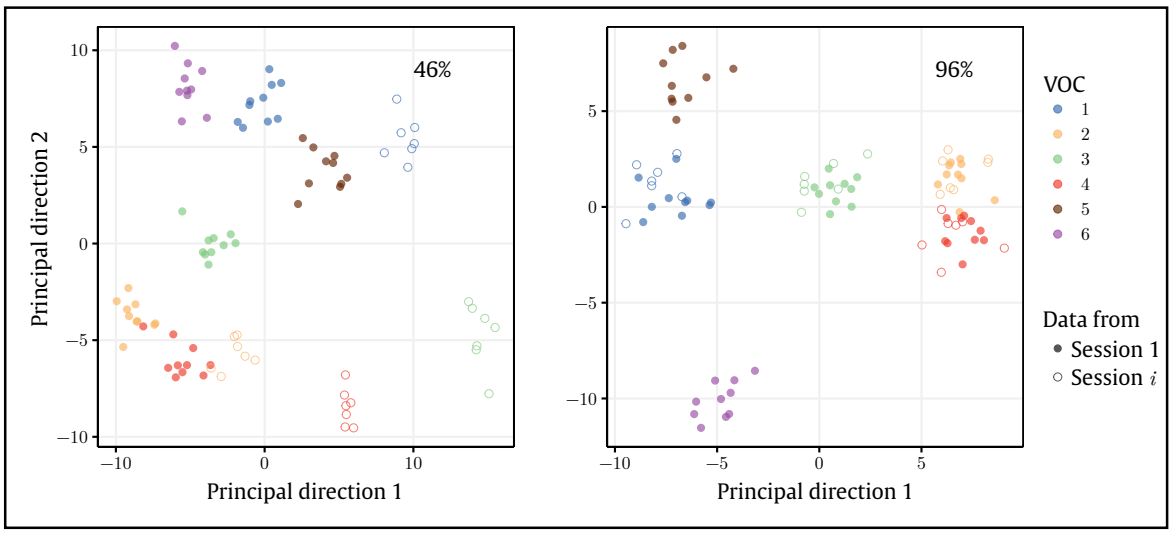
[4] Due to the high computational time required for CMA-ES (the optimization method used by Di Carlo et al.) and due to the high number of simulations, we preferred to use a quasi-Newton method which estimates $\boldsymbol{M}$ much faster while giving similar results as [DC+11]. In particular, we used BFGS method from [Bro70; Fle70; Gol70; Sha70] with the R package optim. To give some numerical values, BFGS method processes a single session in few seconds while CMA-ES can require up to 20 minutes.

corresponds to a total of 28 time windows and for the Discontinuous drift to a single time window. Each time window is considered as a measurement session of $9 \times \gamma$ samples (Session i) which needs to be drift corrected. In other words, for each session and for our method, we estimate a set of parameters $\boldsymbol{\alpha}_i \in \mathbb{R}^R$, $\boldsymbol{d}_i \in \mathbb{R}^P$ and $\boldsymbol{\ell}_i \in \mathbb{R}^{\Delta_t+1}$. For the method of Di Carlo et al., for each session i, we estimate another set of parameters $\boldsymbol{M}_i \in \mathbb{R}^{P\times P}$ and $\boldsymbol{\ell}_i \in \mathbb{R}^{\Delta_t+1}$. $\boldsymbol{M}_i$ is used as initialization for the next session.

Two random realizations are reported in Figure 4.10 and classification results are in Table 4.3.



(a) Continuous Drift



(b) Discontinuous Drift

Fig. 4.10.: Two different random realizations for the Continuous drift (Figure (a), $\gamma = 3$) and the Discontinuous drift (Figure (b), $\gamma = 4$). The indicated scores correspond to classification rates. Left figures correspond to raw data (without any correction) and right figures correspond to data corrected using EMC^2.

γ	Chance level	Continuous drift				Discontinuous drift			
		Raw	PCA-CC	Di Carlo et al.	EMC^2	Raw	PCA-CC	Di Carlo et al.	EMC^2
6	17	64±10	62±16	80±15	98±2	50±19	82±17	51±21	99±3
5	20	65±13	60±18	85±12	98±2	52±22	83±15	50±24	98±5
4	25	65±15	61±21	89±14	97±2	49±24	81±20	51±26	99±3
3	33	65±17	60±23	93±10	97±3	49±27	83±20	53±30	98±6
2	50	64±21	59±27	95±11	89±9	49±32	85±23	50±33	90±26

Tab. 4.3.: Classification results ± standard deviation, over the 100 simulations.

EMC^2 clearly outperforms PCA-CC even if the latter requires much more information. We can draw this conclusion for both Continuous and Discontinuous drift. We notice that both methods are stable regarding γ (the number of VOCs in the future sessions), whatever the performance. It is noteworthy for EMC^2 and predictable for PCA-CC. The parameter γ mostly influences the variance of the classification rate: the lower γ is, the more one misclassification is influencing the score. The performance of PCA-CC is surprisingly quite poor in our simulation settings, in particular for Continuous drift for which the results of PCA-CC are even below those of Raw method. We explain that by the low number of chemical sensors that we simulated ($P = 4$) in addition to the complexity of the drift. Even if Artursson et al. [Art+00] claimed that PCA-CC can compensate for a non-linear drift by removing more and more components, we noticed that this method tends to remove discriminative directions. In fact, due to the low number of chemical sensors, it is more likely that the Principal directions of $\boldsymbol{X}_{\text{cal}}$ (the matrix containing the calibrating samples) carry information about class discrimination. To highlight this remark, we realized other simulations with the same settings, but we increased P to 6. In this case, the average score of PCA-CC for Continuous drift reached 82 % (averaged over γ) and 93% for Discontinuous drift.

EMC^2 also outperforms the method of Di Carlo et al., both for Continuous and Discontinuous drift even if the results are even more clear for Discontinuous drift. For Continuous drift, it is interesting to note that the performance of the method of Di Carlo et al. is quite influenced by γ: the lower γ is, the better. We explain that by the fact that their method makes use of the correction matrix $\boldsymbol{M}_i$ of Session i as initialization for Session $i + 1$. So, this method will tend to accumulate errors and errors will likely increase with γ. In fact, the classification rate over time reported in Figure 4.11 tends to considerably decrease (starting from 100%). EMC^2 does not suffer of this drawback since each session (or time

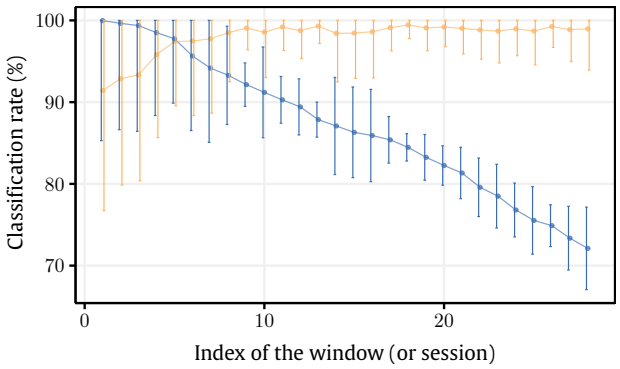


Fig. 4.11.: Classification rate over time for EMC^2 (yellow) and the method of Di Carlo et al. (blue). Error bars correspond to standard deviations over 100 simulations (Continuous drift, $\gamma = 4$).

window) is treated separately. So, if a session is badly corrected, it will not influence the next session. However, Figure 4.11 also shows that EMC^2 behaves badly for the first sessions compared to the method of Di Carlo et al. In fact, since the first sessions may not suffer from drift, the initialization (Algorithm 9) and EMC^2 may find a noisy direction correlated to classes discrimination. This remark is not really a drawback since there is no need for a drift correction method if there is no drift.

4.5 Experimental

In this section, we test Moving Recentering (Algorithm 7) and EMC^2 (Algorithm 8) for the drift correction of a real data set. This real data set was acquired using the robotic platform introduced in Chapter 2, Section 2.3.1. The data set includes several measurement sessions of 3 VOCs acquired over a period of 9 months.

4.5.1 Introduction of the data set

Three gas sources of pure VOCs (Geranyl acetate, Citral and β-Pinene) are disseminated in an open environment. This environment is scanned by a robot which carries the eNose. By following a predefined path, the robot goes over the gas sources and the eNose detects them. To have a better idea of the setup and of the experiments, we encourage the reader to watch a video[5] that I have made during the thesis and available at :

`http://simonbarthelme.eu/personal-website/videos/robot12VOC.mp4`

In this video, a larger setup is used but the idea remains the same. The path is repeated multiple times to obtain a substantial data set at each measurement session. The setup and a raw response (after the baseline subtraction) of a chemical sensor are reported in Figure 4.12. The responses of the chemical sensors always look like peaks due to the very short time injection ($\sim$ 1 sec). A segmentation procedure is applied to extract these peaks. This procedure is described in Appendix A.

After the segmentation, the position in time of each peak is known. Then, we need to extract some features from these peaks. At this stage, we have several options: extract slopes from adsorption and desorption, fit a model and extract the parameters, use all the points of the peak, etc... Here, we choose integration. This feature was in fact the most robust[6]. So, each peak is integrated for each chemical sensor. The eNose is equipped with $P = 63$ chemical sensors, which corresponds to $M = 26$ sensing materials replicated 2 or 3 times on the surface. So, each peak gives a signature in $\mathbb{R}^{63}$. A session usually lasts for few hours so some VOCs run out at the end of the experiment and injected concentration decreases. To get rid of the

[5] If the link does not work, try this one:
`https://drive.google.com/file/d/1AddmCYFwQcHtG1aTksn8zyC-0e7CDzbU/view`

[6] In Chapter 6, we develop another way to process signals by using directly each time point. It bypasses the need of extracting a feature but we observed a little decrease in classification rate by doing so (compared to the feature based on integration). Anyway, these two works (Chapter 4 and Chapter 6) must be combined in a further study.

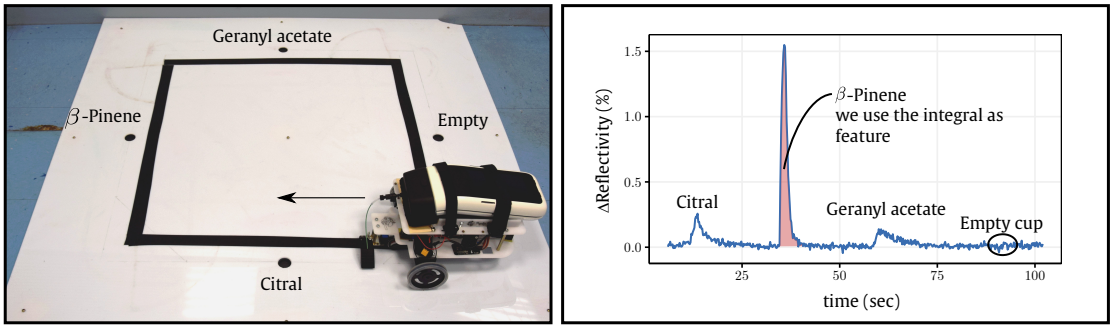


Fig. 4.12.: One of the two robotic setups used for the generation of the real data set (left) and the response of one chemical sensor after one lap (right).

variations in concentration, we normalize each signature individually by dividing it by its ℓ_2-norm.

One lap gives one signature for each one of the $R = 3$ VOCs, so after a certain number of laps we have a data matrix $\boldsymbol{X} \in \mathbb{R}^{N \times P}$. $\boldsymbol{X}$ corresponds to only one measurement session lasting for few hours. Since we perform several measurement sessions over time, we note $\boldsymbol{X}_i \in \mathbb{R}^{N_i \times P}$ the data acquired during Session i.

In total, we have acquired 15 different sessions at different time over 9 months, corresponding to a total of ∼2000 samples. We emphasize that both the number of samples and the total duration are high compared to other data sets available in the literature. The measurement sessions differ on several experimental parameters: the amount of liquid solution left in the gas sources, the environmental conditions (temperature, humidity), the running order, the robot speed and the number of laps. Some of these experimental parameters and the recording dates are reported in Table 4.4.

To add another difficulty, all the sessions have not been generated with the same setup. Most of the sessions have been generated with a setup containing only the 3 VOCs of interest (see Figure 4.12). However, some of the sessions have been generated with a larger similar setup containing 12 different VOCs in 24 different cups, including the studied VOCs (see Figure 2.3 in Chapter 2). For these sessions, we extracted from the data the 3 VOCs of interest. In Table 4.4, the column N_{VOC} indicates the number of VOCs actually present in the session. The combination of all the variations which can exist from one session to another makes this data set really challenging and, with no equivalent in the literature.

In Table 4.4, we also report some classification results. They correspond to the ability of the eNose to tell the 3 VOCs apart when the training and the testing are performed in the same session. These scores have been obtained after a dimensionality reduction with PCA using 5 components and after a 5-fold cross validation using a linear SVM. They highlight that the performance of the eNose regarding these 3 VOCs is stable over time. However, if one wants to use only Session 1 to predict the future sessions, we will see later that a drift between the sessions appeared.

Session	Date	N_i	N_{VOC}	Temp. (°C)	RH (%)	AH (mg/m^3)	SVM (%)
1	Jul 6, 2018	110	3	NA	NA	NA	98 $\pm$ 1
2	Sep 14, 2018	112	3	NA	NA	NA	94 $\pm$ 2
3	Sep 21, 2018	93	3	[31.6, 34.5]	[27.0, 36.0]	[102.9, 119.8]	90 $\pm$ 1
4	Oct 4, 2018	165	3	[25.1, 29.6]	[30.1, 39.0]	[88.3, 93.9]	93 $\pm$ 1
5	Oct 16, 2018	71	3	[28.0, 28.9]	[34.6, 37.5]	[98.5, 103.1]	92 $\pm$ 1
6	Nov 9, 2018	142	3	[23.3, 28.2]	[31.4, 37.5]	[76.9, 87.5]	81 $\pm$ 1
7	Nov 23, 2018	132	12	[20.4, 27.6]	[26.6, 40.0]	[70.3, 74.7]	89 $\pm$ 1
8	Dec 4, 2018	148	3	[23.5, 27.1]	[34.7, 40.3]	[84.8, 94.4]	91 $\pm$ 1
9	Dec 14, 2018	173	3	[24.4, 27.8]	[18.6, 23.7]	[49.8, 56.5]	91 $\pm$ 1
10	Dec 20, 2018	161	12	[26.2, 27.5]	[24.1, 29.3]	[61.3, 73]	76 $\pm$ 1
11	Jan 9, 2019	75	12	[25.1, 27.5]	[20.3, 24.7]	[53.6, 59.0]	71 $\pm$ 3
12	Jan 11, 2019	71	3	[25.6, 27.3]	[14.2, 16]	[36.6, 38.1]	89 $\pm$ 2
13	Feb 6, 2019	164	3	[24.2, 27.5]	[17.2, 21.3]	[45.3, 48.0]	91 $\pm$ 1
14	Feb 13, 2019	73	12	[24.4, 27.0]	[17.6, 23.3]	[44.8, 53.3]	87 $\pm$ 1
15	Mar 21, 2019	256	3	[24.4, 27.2]	[18.1, 22.7]	[46.4, 50.8]	94 $\pm$ 0
	9 months	1934		[20.4, 34.5]	[14.2, 40.3]	[36.6, 119.8]	[78, 100]

Tab. 4.4.: Information about the data sets studied in this section. *NA* values are due to the absence of the sensors at the recording date. RH and AH respectively stand for Relative Humidity and Absolute Humidity. SVM column corresponds to the cross-validated classification rate (repeated 10 times) when we train and test on the same session (mean $\pm$ standard deviation).

4.5.2 Results

Is there any drift in data?

Recall that the data is composed of 15 sessions over 9 months. In practice, one would want to learn from Session 1 and then use this knowledge for the future sessions. Figure 4.13 represents the projection of all the sessions onto the 2 first Principal directions of Session 1. Clearly, a drift appears and tends to stabilize at the end. Classification rates have been obtained after reduction by PCA (using only Session 1 and taking the 5 first Principal directions) and training on Session 1. They considerably decrease after Session 2. Even if there is a visible drift between Session 1 and Session 2, the classifier still succeeds in well

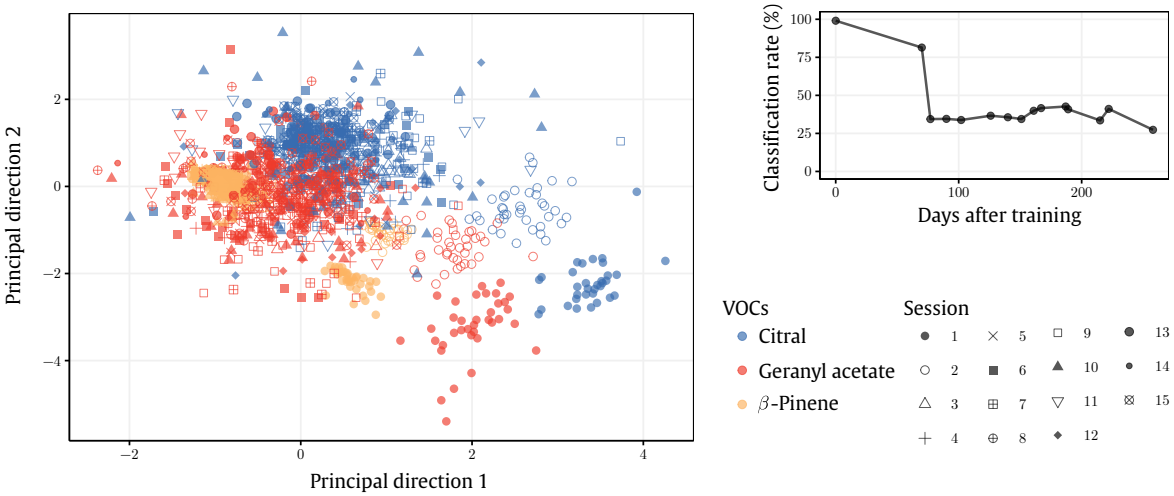


Fig. 4.13.: To the left, we report the projection of all the sessions onto the 2 first Principal directions of Session 1 (the map is centered and scaled for easier visualization). To the right, we report classification rates by training on Session 1.

predicting labels of Session 2. After Session 2, the drift becomes too important and two clusters out of 3 are misclassified. Consequently, results call for drift compensation methods.

How can we compensate for the drift?

To compensate for the drift, we test all the methods described in Section 4.2. We recall that they assume different scenarios and are clearly not on the same level of difficulty:

- **Calibrant Scenario** (CS). Methods assume the existence of calibrating samples during each session. We test PCA-CC (Algorithm 2) and PLS-CC (Algorithm 3) from Artursson et al. [Art+00].
- **Multi-Session Scenario** (MSS). Methods assume that several sessions have been labelled. Here, we assume that the first 2 sessions have been labelled. We test CPCA-CC from Ziyatdinov et al. [Ziy+10] (Algorithm 4) and OSC from Padilla et al. [Pad+10] (Algorithm 5).
- **Blind Scenario** (BS). Methods assume nothing else than the labels from Session 1. We test Di Carlo's method [DC+11] (Algorithm 6). We also test the 2 proposed methods: MR (Algorithm 7) and EMC^2(Algorithm 8). EMC^2 is initialized with Algorithm 9.

Before testing each method, a preprocessing step is required: dimensionality reduction. Indeed, the eNose used for this data set contains $P = 63$ chemical sensors, whereas sometimes the number of samples in a session drops to 71 (*cf* Table 4.4). For classification purpose, the number of chemical sensors is much too high and algorithms could suffer from the effects of the **curse of dimensionality**. For mitigating the effects of the curse of dimensionality, an intuitive method is to simply reduce dimensionality. A first way to do that is to average the replicas of each sensing material: the number of dimensions drops then to $M = 26$ (the number of sensing materials). Each session is now in sensing materials space and we note the data from Session i, $\mathring{\boldsymbol{X}}_i \in \mathbb{R}^{N_i \times M}$.

However, there is still an open question: should we have to compensate for the drift then reduce dimensionality (option 1) or reduce dimensionality then compensate for the drift (option 2)? Both options were tested for all the methods and, except MR, they all performed better with option 1 than with option 2. Since each session lasts for only few hours, MR corresponds to a simple recentering in this case. Therefore projecting onto a lower dimensional subspace or not does not make any difference for MR. So, we choose option 1 (except for Di Carlo's method), meaning that we first compensate for the drift then we reduce dimensionality.

In the Calibrant Scenario, we apply PCA-CC and PLS-CC by using Citral as calibrant and by removing only one component ($K = 1$). For PLS-CC, the recording time is used as dependent variable $\boldsymbol{z}_{\text{cal}}$.

In the Multi-Session Scenario, we apply OSC and CPCA-CC by using the first 2 sessions and by removing also one component. However, only the first session is used for reducing the dimension and training.

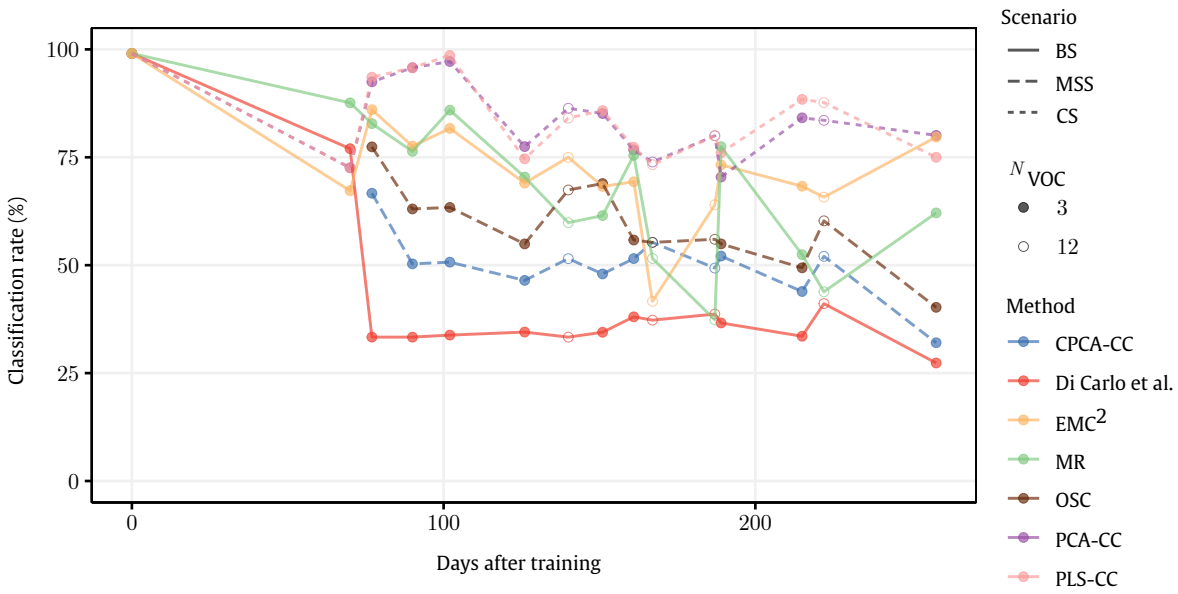


Fig. 4.14.: Results for all the methods used. Only Session 1 is used for reducing dimensions and training (linear SVM).

In the Blind Scenario, Di Carlo's method requires the estimation of a correction matrix of dimension $M \times M$ whereas the number of samples is always lower. So, for this method only, we first reduce and then correct (*i.e.* we use option 2). For the method that we proposed, EMC^2, the dimension can also be an issue, especially for the estimation of the covariance matrices. However, as explained in Section 4.3.2, a variant can be easily implemented by taking diagonal covariance matrices. In the following, we use this variant of the algorithm and the R covariances matrices are assumed to be diagonal (we simply set to 0 the off-diagonal terms).

Classification scores are obtained after dimensionality reduction (based on PCA and on Session 1) and after training a linear SVM using Session 1 (parameters of the SVM are tuned using a cross-validation procedure). To allow a fair comparison between methods, the labels of the corrected data (for each method) are predicted using this linear SVM trained on Session 1 (*i.e.* we do not use the labels which can be automatically returned by some methods).

Method	Scenario	Score (%)
PLS-CC	CS	84.1
PCA-CC	CS	83.7
EMC^2	BS	72.4
MR	BS	68.2
OSC	MSS	59.0
CPCA-CC	MSS	50.0
Raw	-	43.8
Di Carlo	BS	42.1

Tab. 4.5.: Ranked averaged scores over sessions.

In Figure 4.14, we report the classification scores for all the methods used. Only Session 1 was used for reducing dimensions and training a linear SVM. In Table 4.5, we report the averaged classification scores over sessions (excluding Session 1).

Clearly, methods using calibrating samples outperfom the others. They are followed by the two methods that we have proposed: MR and EMC^2.

This shows that the drift model is appropriate despite its simplicity. Methods based on several labelled sessions perform poorly. This poor performance can be attributed to the small number of training sessions that we have used. The 2 first sessions may not contain enough drift information about the subsequent sessions (although Session 2 is clearly drifting as shown by Figure 4.13). Performance may be increased by considering more training sessions. However, this result is interesting since it highlights their main drawback: in practice, nobody knows how many training sessions are required for the future. The last method used, namely Di Carlo's method, has the same performance than Raw method, the method which does not apply any correction. This reflects one of the points that we have raised in Section 4.2.3 concerning this method: it cannot adapt to a Discontinuous drift. In Figure 4.13, the discontinuous nature of the drift is clearly perceptible in the data. This method would likely perform better if sessions had been less spaced in time. However, the sampling of the sessions is often not controlled in the field and methods have to cope with both Continuous and Discontinuous drift.

What would be the performance if Session 1 measures R VOCs while future sessions are measuring $\gamma < R$ VOCs?

Previous results are not perfect since scores are still not at the level of intra-session scores (when we test and train on the same session) that we reported in Table 4.4 (SVM column). However, most of the correction methods enable better performance than Raw method (the method which does not correct anything). To make the task harder, we can consider a scenario which can occur in practical applications: Session 1 measures R VOCs while future sessions are measuring a smaller number of VOCs, say γ. The identity of these γ VOCs is obviously unknown, which makes the task of drift correction even more tricky.

Unfortunately, in our data set, all the sessions measure the same number $R = 3$ of VOCs. However, we can artificially remove from the subsequent sessions one VOC out of the R VOCs, leading to $\gamma = 2$ VOCs. Only the first session (and the 2 first sessions for methods considering the MSS) keeps $R = 3$ VOCs for the training. Drift correction methods are then applied according to the same pipelines. For methods requiring calibrating samples, one VOC out of the $\gamma = 2$ VOCs is randomly selected as the calibrant for all the sessions. For selecting which VOC is removed from all the testing sessions, the solution is simple: in separated experiments, each one of the $R = 3$ VOCs is selected in turn.

Table 4.6 reports the classification results, averaged over sessions (excluding Session 1). The ranking is more or less unchanged compared to Table 4.5. Methods considering the CS still outperform the others. However, $\gamma = 2$ which means that 50% of the information is known by these methods (information from calibrating samples), so the significance of their performance must be moderated. More surprisingly, the proposed method EMC^2 performs

Method	Scenario	VOC out			Average (%)
		Citral	Geranyl acetate	β-Pinene	
PLS-CC	CS	84.9	91.4	76.2	84.2
PCA-CC	CS	83	92.3	75.7	83.7
EMC^2	BS	70	58.8	68.7	65.8
OSC	MSS	58.8	80.1	36.9	58.6
CPCA-CC	MSS	55.7	76.2	29.8	53.9
MR	BS	40.6	45	54.4	46.7
Raw		54.9	56	7.6	39.5
DiCarlo	BS	55.5	50.5	0.1	35.3

Tab. 4.6.: Average scores over sessions when one VOC is removed from the subsequent sessions. Results are ranked according to the global average (over sessions and VOC out).

fairly well since, recall that nothing else than Session 1 is known by this method. The other proposed method, MR, is now performing poorly. But this result is expected: with $\gamma = 2$ VOCs, we are no longer considering the scenario for which MR can work (Equal Frequency of Occurrence Scenario). Concerning the other methods (CPCA-CC, OSC, Di Carlo's method), their global performance is poor which is mainly due to the data set where β-Pinene is removed. In fact, most of the good classifications of these methods come from this cluster, by removing it, we remove the good classifications.

4.6 Conclusion and future works

Drift is a fundamental issue affecting chemical sensors and limiting the reproducibility of electronic noses over time. Factors contributing to the drift are numerous and come from phenomena which cannot be easily modelled and which are often (or rather always) concomitant in practical applications.

In this chapter, we have introduced a comprehensive data set composed of 15 different sessions acquired over 9 months in our lab and under uncontrolled conditions. This data set highlighted that the eNose of Aryballe suffers also from this issue.

To compensate for the drift, several methods of the literature have been listed and detailed and we have classified them according to three possible scenarios: methods which are assuming calibrating samples (Calibrant Scenario), methods which are assuming several labelled sessions (Multi-Session Scenario) and methods which are assuming nothing else than the labels from the first measurement session (Blind Scenario). The last group of methods is by far the most challenging and the least explored in the literature. For the Blind Scenario, we have proposed two drift correction methods based on a common drift model, namely Moving Recentering (MR, Algorithm 7) and Expectation-Maximization Component Correction (EMC^2, Algorithm 8).

EMC^2 has been successfully validated on an artificial data set involving a non-linear drift. MR and EMC^2 have been then tested on the real data set that we have generated and which is affected by a significant drift between the first session and the subsequent sessions. They

have also been compared to the methods from the literature, whatever the group of methods. We have also studied the realistic case where the number γ of VOCs in the future sessions is smaller than the number R of VOCs in the first session (used for the training).

Most of the methods allowed the drift to be partly compensated. Calibrant-based methods achieved the best performance, as long as they have calibrating samples, which can be time-consuming in practice. The two proposed methods performed fairly well given the difficulty of the task, outperforming some state-of-the-art methods. MR can work only if class proportions remain stable over time so its performance decreased when we have considered the case $\gamma < R$. The performance of EMC^2 was satisfactory for both cases ($\gamma = R$ and $\gamma < R$). All the other methods were less efficient according to the data set and to the settings that we have used.

What are the perspectives?

Drift is still an open issue. First, the methods have compensated only partly for the drift so there is still some work for improving the results. The drift model seems appropriate but the solution proposed with EMC^2 is also quite dependent on a good initialization. Other algorithmic methods could be derived from the same model and may be less dependent on the initialization. Besides, we have considered the drift for a classification task. It is still unclear what would give the correction if gas quantification is required (namely the estimation of concentration).

In this chapter, we have been only interested in intra-instrument drift. However, another fundamental issue is inter-instrument drift. This drift may appear due to the weak reproducibility of the chemical sensors. The solutions proposed may generalize to inter-instrument drift but this calls for data sets. Besides, two instruments may also not have the same sensing materials, making the generalization even more difficult.

Stereoselectivity

5

Contents

The results of this chapter have been published in a journal paper:

[hal-02534216] P. Maho, C. Herrier, T. Livache, G. Rolland, P. Comon, and S. Barthelmé
"*Reliable chiral recognition with an optoelectronic nose*"
In: *Biosensors and Bioelectronics* 159 (Jul. 2020)
10.1016/j.bios.2020.112183

5.1 Introduction

What is stereoselectivity?

To define stereoselectivity, we need first to define what **chirality** is. The word chirality is derived from the ancient Greek χειρ ("kheir") which can be translated as hand in English. Formally, chirality is the property of an object to be non-superimposable to its mirror-image. An object having this property is said **chiral** while if it does not have this property it is said **achiral** (the object and its mirror-image can be superimposed). The etymology takes then on its full meaning: hand is a chiral object (see Figure 5.1[1]). What is the mirror-image of the right hand? Left hand. Using translations and rotations, can we superimpose left and right hands in 3D space? Of course not. For a chiral object, there are then two possible forms, left or right. These two forms are called **enantiomers**. Several conventions are used to differentiate them, the convention that we used here is (R) and (S). So, for a chiral object, there will be always a (R)-enantiomer and a (S)-enantiomer. Finally, a technique is said **stereoselective** if it can differentiate between (R) and (S) forms. For instance, a left-handed glove is stereoselective.

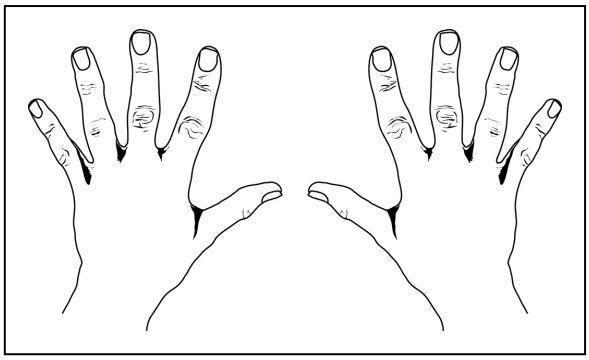


Fig. 5.1.: Left and right hands show chirality at work.

Nature is chiral.

In chemistry, molecules can be chiral. The two forms of a chiral molecule have a lot in common, in particular they have almost identical physico-chemical properties (*e.g.* mass, boiling point, solubility, volatility, etc...). In fact, it does not exist more similar molecules than 2 enantiomers of a chiral molecule. However, they are not exactly the same molecule so there exists two ways for telling two enantiomers apart. First, they do not interact in the same way with polarized light: one enantiomer deviates light to the right and the other to the left. Second, they do not react in the same way with a chiral molecule (but they do have the same reaction with an achiral molecule) [Ben06]. The second principle is at the core of the stereoselectivity of many biological systems.

[1] Figure taken from the website https://clipartart.com/

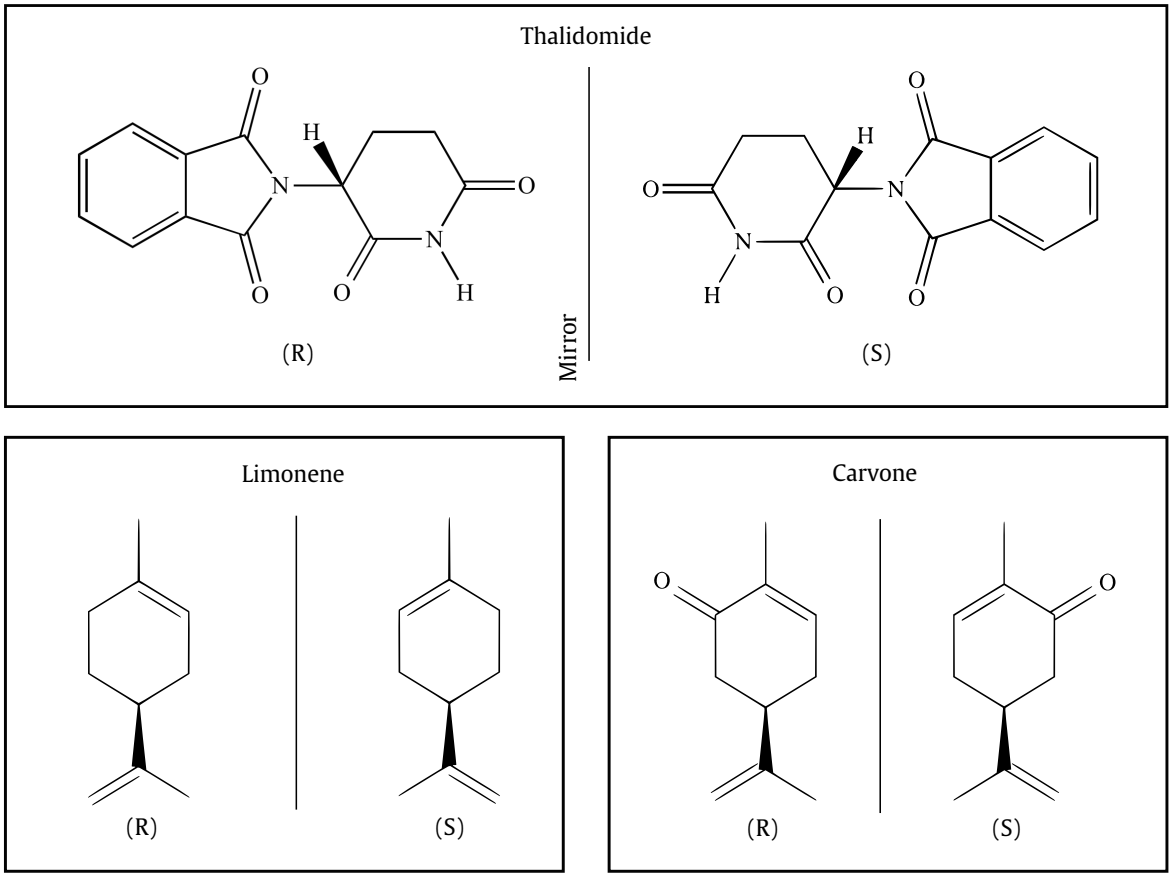


Fig. 5.2.: Shorthand formulas of some chiral molecules. Limonene and Carvone will be further studied.

In Nature, the building blocks of life, *i.e.* amino acids and sugars, are chiral. A surprising and still mysterious fact is that almost all amino acids (out of 22) are left-handed enantiomers [Chy97]. Amino acids are the structural unit of peptides and proteins, which makes them chiral too. The chiral property of proteins explains why there are so many biological processes which are stereoselective.

For instance, human nose is stereoselective [Ben06]. It has been proven that the differentiation between two enantiomers is feasible for some chemical substances. For instance, (R)-Carvone smells spearmint whereas (S)-Carvone smells caraway [RH71]. Another example is Limonene: (R)-Limonene smells orange whereas (S)-Limonene smells lemon [FM71]. Shorhand formulas[2] of Carvone and Limonene are reported in Figure 5.2.

Stereoselectivity is also a crucial concept in pharmacology. In fact, a large part of drugs are chiral and one of the two enantiomers is often preferred since their biological activity can be stereoselective. In some cases, the use of the wrong enantiomer can lead to inactive substance but also to side effects [NHPH06]. An example is the Thalidomide tragedy [KS11]. In the late 1950s, Thalidomide has been marketed as a sedative and was used in the treatment of nausea

[2]Shorthand formula: it represents the skeletal structure of an organic compound as a series of atoms bonded together and forms the essential structure of the compound. The labels C (Carbon) and H (Hydrogen) are generally not represented and only the main bonds (C-C or between C and another atom) are represented to lighten the structure.

in pregnant women. Few years later, several thousands of babies were born with missing or abnormal arms, legs, hands or feet. Thalidomide was eventually banned in most countries in 1961. Thalidomide is a chiral molecule and was in fact marketed as a **racemic mixture** (equal proportion of (R) and (S)), so the drug was composed of both (R)-Thalidomide and (S)-Thalidomide. Later, researchers discovered that (R)-Thalidomide is actually a sedative while (S)-Thalidomide can harm a foetus. Thalidomide episode led to a law reinforcement and changed the way drugs are tested [Var15]. However, this tragedy could not have been avoided since some human enzyme can transform (R)-Thalidomide into (S)-Thalidomide [Var15]. So, even a drug only composed of (R)-Thalidomide would probably lead to the same disaster. Shorhand formulas of Thalidomide are reported in Figure 5.2.

In conclusion, stereoselectivity seems to be straightforward for Nature.

What about man's machines?

In the laboratory, the differentiation between two enantiomers is quite challenging and is considered as the most difficult task regarding molecular recognition [Tro14].

Analytical methods. In Section 1.1.2, we cited Gas Chromatography (GC) as the gold standard for separating a gas mixture into its individual components. Separating two enantiomers is so complicated that even this analytical method fails. To allow GC to be stereoselective, that requires to use a chiral stationary phase (the stationary phase corresponds to the molecule covering the internal surface of a GC column), leading to the so-called chiral-GC [Sch02]. Another method is called derivatization [Sch02]. The idea is to use a chemical reaction creating new molecules which are simpler to separate. For separating a racemic mixture (equal proportion of the 2 enantiomers), we introduce a pure chiral reagent which breaks the chirality by creating two new molecules, called diastereoisomers. Two diastereoisomers are quite similar molecules but more different than two enantiomers. In fact, one diastereoisomer is not the direct mirror-image of the other. Consequently, they will not share the same physico-chemical properties, making their differentiation easier. In fact, they can be separated by a classical GC. So, by injecting in a GC the new mixture obtained after derivatization, we can separate the two enantiomers. These methods suffer from the same drawbacks as those previously cited in Chapter 1 for GC (expensive, requires qualified personnel, cannot be used on a large scale, etc...).

During the last decade, the use of chemical sensors for this task, instead of analytical methods, aroused interest and new technologies have been developed [Tro14].

Highly-specific sensors. The most popular method is to build a specific device to the targeted chiral molecule. One good example of such chiral devices is sensors based on Molecularly Imprinted Polymers (MIPs). MIPs artificially create recognition sites for a given VOC (for instance, (R)-enantiomer is used as the template) which we want to detect. First, some molecules, called functional monomers, bind to the template: this is the pre-arrangement step [TP15]. Second, the complex formed by the template and the monomers interacts

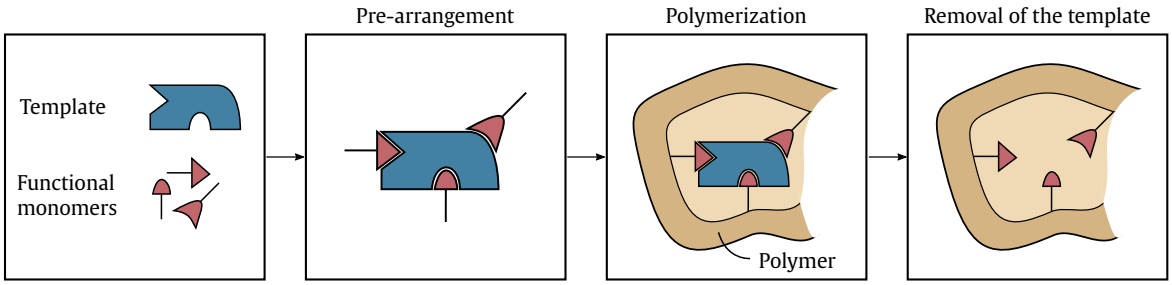


Fig. 5.3.: Principle of Molecularly Imprinted Polymers, inspired from [TP15]. First, functional monomers bind to the template (pre-arrangement). Second, the complex formed by the template and the monomers interacts with another polymer (polymerization). Finally, the template is removed leaving cavities inside the polymer. These cavities have the same shape and size as the template.

with another polymer (*i.e.* a chain formed by the repetition of single molecule): this is the polymerization step (*i.e.* small molecules interacts to form a larger molecule) [TP15]. Finally, the template is removed leaving cavities inside the polymer due to the initial monomers. These cavities have the same shape and size as the template. So, the resulting sensor will have a better affinity to the (R) form than (S) form (see [TP15] for an example of a stereoselective sensor based on MIPs). In Figure 5.3, we report the principle of a MIP-based sensor. This kind of sensors will be highly-specific to the targeted chiral molecule but will not be useful for other analyses.

Electronic noses. Compared to the previous group, electronic noses rely on non-specific chemical sensors which can be used for a broad variety of applications. These instruments are expected to be universal as their biological counterpart. So, some authors have tried to show the stereoselectivity of eNoses, at least for two technologies. First, Brudzewski et al. [Bru+07] explored the stereoselectivity of MOS sensors, using 7 elements (TGS sensors from Figaro Engineering Inc.) and testing Carvone and Limonene. Based only on 10 samples, they concluded to a perfect discrimination between the two enantiomers of both chiral molecules. Their conclusion was motivated by the crystalline nature of the MOS surface. Kybert et al. [Kyb+13] reported the use of carbon nanotubes coated with single-stranded DNA for differentiating the enantiomers of Limonene and α-Pinene. They motivated this result by the chiral nature of DNA. They suggested that their sensors could be incorporated in an electronic nose system.

What is the goal of this chapter?

The goal is simple: show the stereoselectivity of Aryballe's eNose.

The motivation is also simple: multiple sensing materials used by the instrument of Aryballe are peptides. We have seen earlier in this introduction that, peptides are chiral and enantiomers may react differently in the presence of a chiral sensor.

However, the demonstration is not that simple. First, chirality of the sensors is a necessary condition but not sufficient for stereoselectivity [Ben06]. Since the instrument is not

designed for stereoselectivity, its possible stereoselective behaviour is not obvious at all. Second, differentiating between two enantiomers is the hardest problem regarding molecular recognition [Tro14]. If observed, the differences must be really tiny so these differences must be questioned in relation to possible experimental confounds and statistical reliability.

A naive approach for showing the stereoselectivity of the instrument would be to use the automatic valve described in Chapter 2 and to generate a large data set, thus ensuring the statistical reliability. However, we will see in this chapter (in Section 5.4.1) an experimental confound which is surprising and unexpected. The instrument is so sensitive that it can make a difference between two different vials containing the same compound. We call this unexpected differentiation, the vial differentiation. The difference is tiny but the same applies to the difference between two enantiomers. The reason of the vial differentiation is unclear but the problem is obvious: how can we show that the observed difference between two enantiomers is not only related to the fact that the two enantiomers are necessarily in two different vials? The naive approach is not sufficient and we describe in the following a multi-session approach and an extensive discriminative analysis to alleviate the vial differentiation.

5.2 Experimental: a multi-session approach

5.2.1 Experimental Setup

Chemical compounds. We test the enantiomers of Limonene and Carvone, since the human nose is stereoselective for these 2 compounds [RH71; FM71]. All the products were purchased from the same seller (Sigma-Aldrich): Butanol, mineral oil, (R)-Carvone pure at 99.4%, (S)-Carvone pure at 99.3%, (R)-Limonene pure at 98.4% and (S)-Limonene pure at 97.9%. The purity of each enantiomer comes from the certificate of analysis of Sigma-Aldrich. Due to the volatility of Butanol and to the duration of the experiments, we conducted a volumetric dilution at 25% in mineral oil to avoid exhaustion. This dilution is only conducted on Butanol and the other chemical compounds are left pure and used as received without further purification. The purity degree of the samples is in line with previous studies using the same compounds [Bru+07] (and even better).

Gas sampling system. In this study, we use the automatic gas sampling system described in Section 2.2 and reported in Figure 5.4, see Chapter 2 for details.

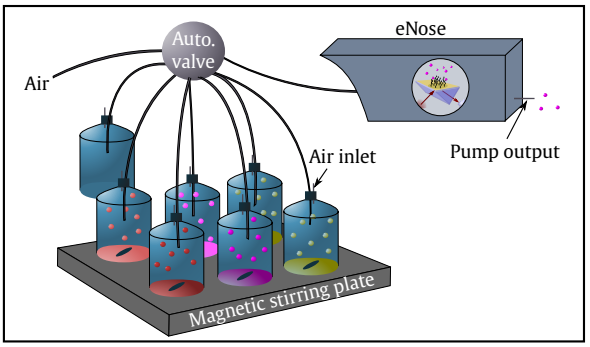


Fig. 5.4.: Automatic valve.

All lines except the reference line are inserted in 50 mL vials, which are sealed with a Teflon-coated silicon septum and filled with a liquid solution of each VOC. The time segmentation used is as follows: 30 sec for the reference gas (*ie* ambient air), 30 sec for VOC injection and 5 min for desorption. Only the

two first phases are recorded and the third one is deliberately long, as a precaution. Consequently, a valve's cycle lasts 42 min, corresponding to 7 measurements.

Optoelectronic nose. Aryballe's eNose is detailed in Section 1.3. The instrument used in this chapter boards $M = 19$ different sensing materials deposited on the surface of the prism. 17 of them are peptides, so are chiral sensors. The two remaining sensing materials are achiral molecules and are used as a control for the study. Each sensing material is repeated 3 or 4 times on the surface, leading to a cross-sensitive chemical sensor array of $P = 59$ elements. The flow of air of the pump is set to 63 mL/min.

5.2.2 Multi-session protocol

Showing the stereoselectivity of an instrument is a difficult task and requires care, especially regarding statistical reliability and experimental confounds. The use of the automatic valve allows the statistical reliability to be improved, especially by generating more than 100 samples/class.

Experimental confounds are reduced by repeating the experiment twice over two long sessions, controlling various factors. Between sessions, all the vials are changed for Session 2 and filled with new products taken from the same stock solutions as in Session 1. We also change the liquid volume between the two sessions, adding another source of variability. The running order is also not the same to exclude any memory effect (*i.e.* artifact-based discrimination could be due to some traces of the previously measured VOCs in the sampling system or on the sensing surface) [DAS03]. Finally, to keep realistic data (mimicking an industrial application), the whole setup is left at room conditions during the two long sessions. Temperature and humidity variations are then expected. The lines allocation and other parameters are reported in Table 5.1.

Session 1 lasted ∼88 hours, providing 125 samples/molecule and Session 2 lasted ∼182 hours for 260 samples/molecule, both run without interruption. Session 1 and Session 2 are separated by a single day and the total duration of the experiments is ∼13 days. So in total, the evaluation of the stereoselectivity for a given pair of enantiomers is based on ∼800 samples. Humidity and temperature variations are reported in Figure 5.5, showing the day/night cycle.

	Session 1		Session 2	
Line	Molecule	V (mL)	Molecule	V (mL)
①	Air (baseline)	-	Air (baseline)	-
②	(R)-Limonene	0.4	(S)-Carvone	0.150
③	Butanol	1.6 (at 25%)	Butanol	1.6 (at 25%)
④	(S)-Limonene	0.4	(R)-Carvone	0.150
⑤	Butanol	1.6 (at 25%)	Butanol	1.6 (at 25%)
⑥	(R)-Carvone	0.4	(S)-Limonene	0.150
⑦	Empty vial	-	Empty vial	-
⑧	(S)-Carvone	0.4	(R)-Limonene	0.150

Tab. 5.1.: Experimental protocol reporting the lines allocation for each session.

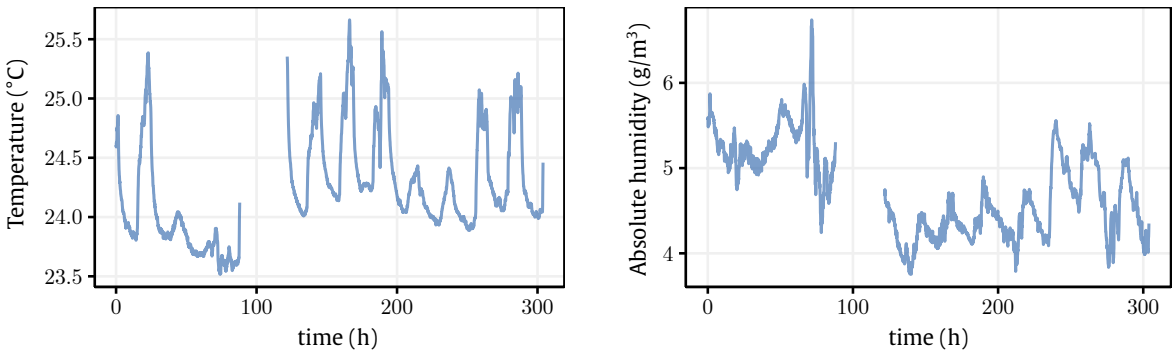


Fig. 5.5.: Temperature and absolute humidity variations over the 2 sessions.

5.2.3 Data processing

Baseline manipulation. For each chemical sensor, baseline value is subtracted (computed by averaging the first 25 sec of signal). This preprocessing helps to remove, at least partially, some drift due to a reference change [MGG12]. In Figure 5.6a, we represent the time series of the $P = 59$ chemical sensors after subtracting the baseline.

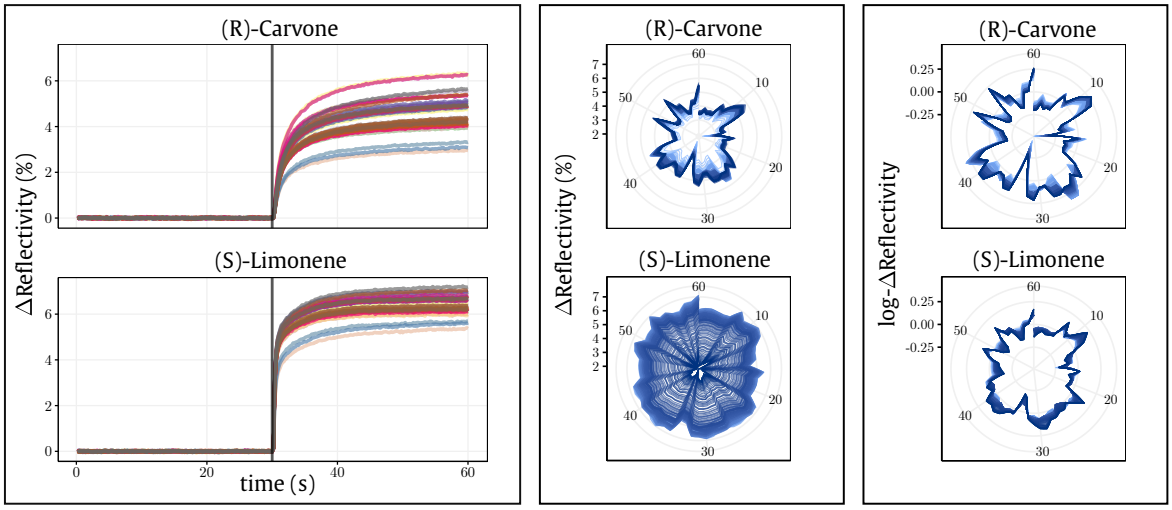
Feature extraction. The most common feature is the steady state value [Yan+15]. This value is chemically justified since a single time series corresponds to the binding reaction between the sensing material and the VOC. This reaction is likely to reach an equilibrium phase, stabilizing the response around a given value. However, due to the schedule we used, some VOCs reach equilibrium while others do not. As an alternative to steady-state values, we propose to integrate the signal over the entire adsorption part. We retain the value of the integral as a feature (*e.g.* the integration over the 30 seconds after the molecule injection). We saw in Chapter 3 (Section 3.2.3) that this feature performs actually well compared to more sophisticated features (for instance, features from a model).

Each chemical sensor provides a feature, and this procedure therefore gives us a vector $\tilde{\boldsymbol{x}}_{rn} \in \mathbb{R}^P$, with P the number of chemical sensors, for the measure of the VOC r during a valve cycle n.

Feature normalization. All measures from Session 2 are represented in polar coordinates in Figure 5.6b. On the same figure, we can easily see the exhaustion occurring over time, especially for Limonene, leading to a change in concentration, which in turn results in a decrease in reflectivity. As this study is interested in qualitative results, variations in concentration cause undesirable variations. To eliminate these variations, we normalize $\tilde{\boldsymbol{x}}_{rn}$ by subtracting the log-average over all chemical sensors, *i.e.*, if subscript p refers to the chemical sensor:

$$x_{prn} = \log(\tilde{x}_{prn}) - \frac{1}{P}\sum_{i=1}^{P} \log(\tilde{x}_{irn}) \tag{5.1}$$

The normalization (5.1) implies that measurements scale linearly with VOC concentration. Although clearly an approximation, it has always worked well in practice. The normalized data



(a) Time series (b) Raw signatures (c) Normalized signatures

Fig. 5.6.: (a) Example of time series obtained for (R)-Carvone and (S)-Limonene. A vertical line indicates the VOC injection. It corresponds to a single measurement, and the baseline has been subtracted. (b) Raw signatures (integration of the adsorption part, over the 30 s after injection in (a)) in polar coordinates (angle for sensor number and radius for amplitude). Color gradation stands for the cycle number (the initial cycle is the lighter, meaning the first measurements), showing for example the gradual exhaustion of Limonene during Session 2. (c) Normalized signatures according to (5.1): repeatability is visibly increased (compared to unnormalized signatures in (b)).

of Session 2 is represented in Figure 5.6c and clearly demonstrates the increased repeatability of the measures, even if some time-related variations remain.

Drift compensation. Remember that sessions are spreading over several days (∼8 days for Session 2) and environmental conditions are not reliably controlled to make data more realistic. In Chapter 4, we deeply discussed a common issue with chemical sensors: their response is not stable over time, we say that they drift. In Figure 5.7, we represent the 2 first principal components for the data from Session 2 and drift is clear. To reuse the terminology that we have introduced in Chapter 4, this data set suffers from a Continuous drift. This Continuous drift seems to be correlated with the variations in temperature and humidity that we reported in Figure 5.5. Several methods have been presented to compensate for the drift. In particular, we proposed two new methods and validated them on both artificial and real data sets. We named these two methods Moving Recentering (MR, Algorithm 7) and Expectation-Maximization Component Correction (EMC^2, Algorithm 8). Here, MR seems appropriate. Indeed, we cannot find two other molecules which are more similar than the two we are looking at, namely two enantiomers (Carvone or Limonene). Consequently, it is very likely that the 2 molecules will be affected by the same amount and direction of drift. This is even more true due to the experimental setup and to the short time spacing between two enantiomeric samples. Besides, class proportions are stable over time. So, we apply MR method on the two sessions separately with $\Delta_t = 5$ hours. Results for Session 2 are reported in Figure 5.7 and show a clear improvement which will be quantitatively assessed in the next section.

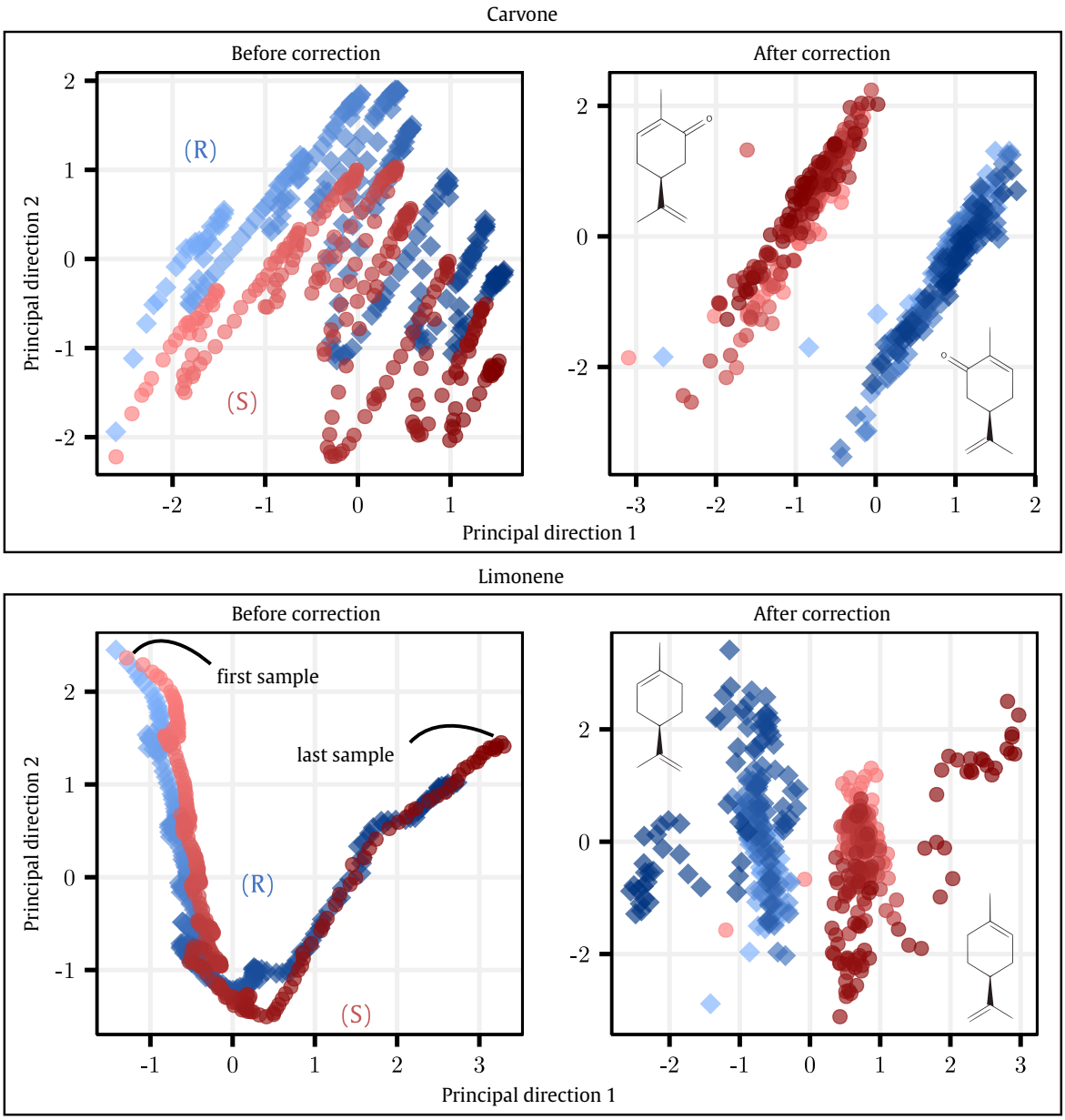


Fig. 5.7.: Projection of the data from Session 2 onto the 2 first principal directions, separately for Carvone (top) and Limonene (bottom). On the right, the drift has been corrected, whereas on the left not. Each point corresponds to a measurement such the one in Figure 5.6a. Cycle number is represented in the color scale (the initial cycle has the lightest colour). The data set used here is from Session 2. The maps have been centered and scaled for easier visualization.

5.3 Stereoselectivity

In this section, we show the stereoselectivity of the optoelectronic nose for Carvone and Limonene. We start by some controls which have been carried out in-between sessions to eliminate potential confounds. Then, we perform a rigid proof of the stereoselectivity based on an intra and inter session analysis.

5.3.1 Controls of the empty setup

A recent paper warned researchers about the underestimated contamination brought by magnetic stir bars, which are often assumed clean after intensive washing [Pen+19], but may retain traces of their previous uses. In fact, the same may hold for every material used in chemical experiments. In our case, these contaminations may mean that the differentiation is artifactual, due to traces left in vials or in the tubing.

First, the fluidic system is deliberately made in PolyEtherEtherKetone (PEEK), a material which is assumed to be non-reactive with any VOC. However, any tiny residue stuck in the system could lead to a tiny difference between two lines, resulting a difference between two VOCs. To remove this eventuality, we intensively checked the cleanliness of the whole fluidic system between Session 1 and Session 2. We ran 20 cycles without any vials. Then, we extracted from the signals the same feature as the one used in the study (integration of the adsorption part). Since no vials were present, this value is expected to equal zero, meaning a clean line. This feature was not normalized to avoid any problems with negative values (the signal level was very low, and negative values could arise due to noise).

We tested the discrimination between pairs of lines booked for a pair of enantiomers (line 2 against 4 and line 6 against 8). This test was as follows: first, we reduced dimensionality by taking the 3 first Principal Components, then performed a repeated 3-fold cross validation (repeated 50 times) with a linear Support Vector Machine from which we average the results to get an estimate of the classification rate between 2 empty lines. The results are reported in Figure 5.8 and show no discrimination due to a pollution in the lines. We highlight that the closer to 50% (chance rate) the results are, the better it is.

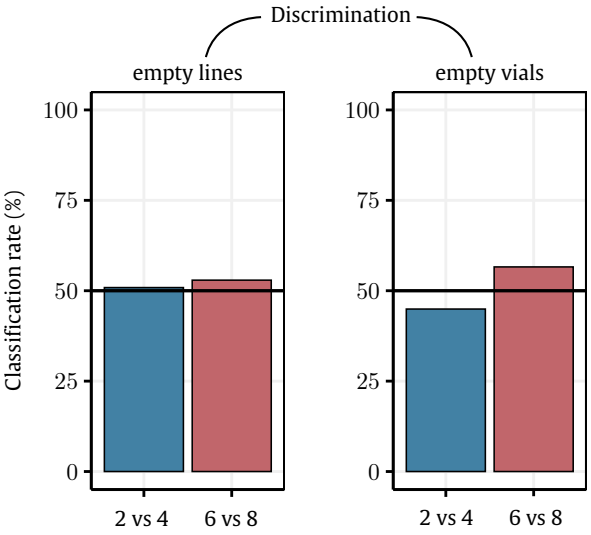


Fig. 5.8.: Discriminative power of the empty setup. Left figure checks if we can find any discrimination based only on gas lines. Right figure checks if we can find any discrimination only based on gas lines, empty vials and magnetic stir bars. A value close to 50% indicates that there is no discrimination. "2 vs 4" means line 2 versus line 4 (in Table 5.1).

Second, the vials were either first-use, or intensively cleaned by rinsing them with Acetone then Ethanol and by drying them for several hours. However, as above, any tiny remaining pollutant in the vials could lead to a tiny discrimination between two vials, meaning between two VOCs. Again, we carried out 20 cycles by placing empty vials and magnetic stir bars. The same discriminative analysis was repeated between pairs of vials booked for a pair of enantiomers. The results are reported in Figure 5.8 and show again no discrimination due to the vials.

5.3.2 Intra-session

We start by proving stereoselectivity when we train and test from the same session. To assess quantitatively both drift compensation and stereoselectivity, we use a chronological validation differing from standard cross-validation (similar to the scheme in [Pad+10]). Standard cross-validation divides the dataset into random pieces (folds), disregarding temporal order. This enables classifiers to "learn the drift", meaning that they find a linear discriminant that is orthogonal to the drift direction. Instead, we split each session into continuous blocks of one day each. Then, a single one-day block is taken for training (which includes both the estimation of the principal components and the fitting of SVM classifier), and all other blocks are used for validation (estimating discrimination performance). This process is repeated for each day and the classification rates are then averaged: this method thus answers the question of how reliable the signatures learned on a single day are, when used on another day (and drift may have occurred in between). The validation scheme is reported in Figure 5.9.

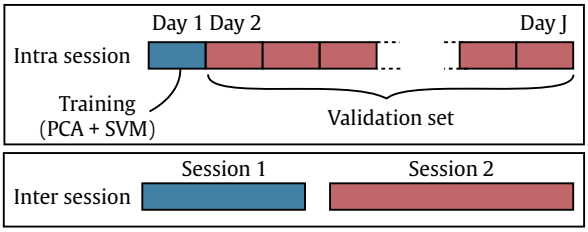


Fig. 5.9.: Validation scheme used for assessing drift correction.

The comparison of results with and without drift compensation is indicated in Table 5.2. A clear improvement can be noted after drift compensation by MR method. Concerning stereoselectivity, classification rates averaged over sessions, reach the almost perfect scores of 99.3% and 99.8% respectively for Carvone and Limonene. These scores clearly indicate that the eNose can learn to discriminate the vial containing the left-hand form from the one containing the right-hand form.

5.3.3 Inter-session

Stereoselectivity with a non-specific device is a very hard discrimination task, which requires a very sensitive instrument. As a result of the instrument's sensitivity, a successful discrimination could be explained by an experimental artifact instead of an actual discrimination between different enantiomers. To exclude this possibility, we carry out an inter-session analysis which echoes the previous one.

	Session	Before correction (%)	After correction (%)
Carvone	1	86.3	98.9
	2	73.8	99.7
Limonene	1	95.7	100
	2	65.3	99.6

Tab. 5.2.: Comparison with and without drift compensation applied session-by-session. The difference of scores between Session 1 and Session 2 before correction can be explained by the difference of duration. Indeed, Session 2 lasts twice the time of Session 1, thus Session 2 includes more drift.

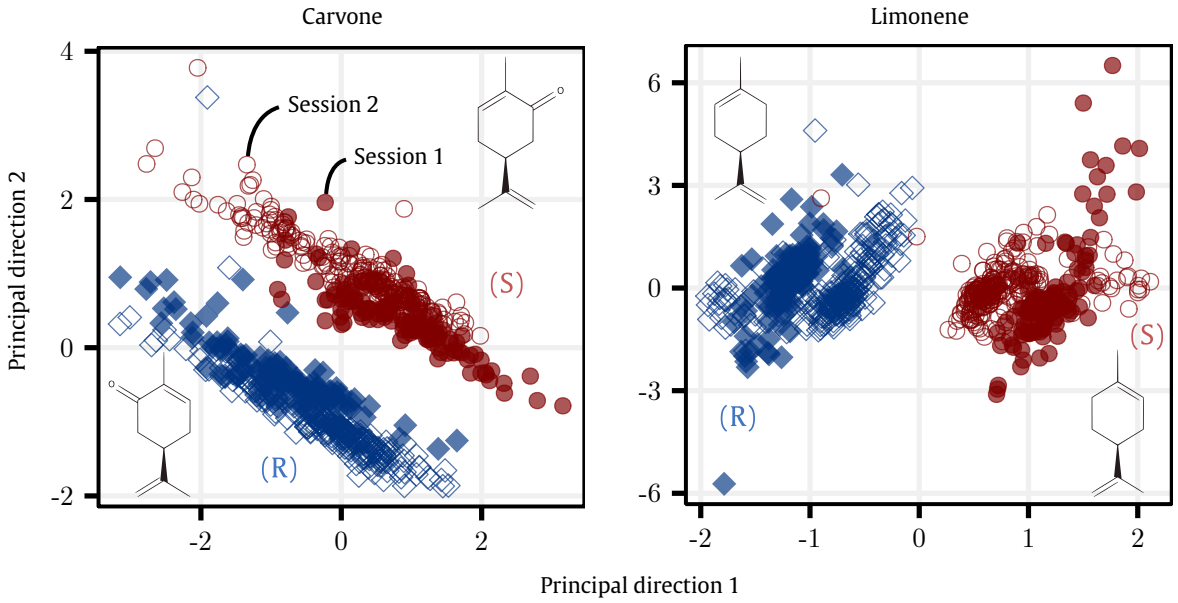


Fig. 5.10.: Multi-session representation. The two data sets have been drift compensated beforehand. The principal components are learnt from the Session 1 and the data from Session 2 are projected in this subspace. Colors indicate the enantiomer. Solid shape is for Session 1 and hollow shape for Session 2. The maps have been centered and scaled for visual consideration.

The two sessions that we have presented so far varied in some of their parameters. As a reminder, there were variations in concentrations, the vials which have been changed and filled with new products (from the same stock solutions), the magnetic stir bars and the running order are not identical from one session to the other. A resulting check is whether one could successfully generalize from one session to the other?

To answer this question, the procedure is as follows: learn to discriminate the 2 mirror molecules using the data from one session and try to classify them in the other session. To clarify, we compute the new representation space (3 first Principal directions) and optimize the SVM from Session 1 and we compute the classification rate obtained for Session 2. This validation methodology is repeated by interchanging the role of each session. The validation scheme is reported in Figure 5.9.

The projection of samples from Session 2 in the representation space generated by the Session 1 is represented in Figure 5.10. Again, numerical results follow the visual impression since the averaged classification rate (across the sessions, each one taken as the train set for the other) reaches respectively 99.9% for Carvone enantiomers and 99.8% for those of Limonene. These results reduce the likelihood of a memory effect or an artifact-based differentiation.

5.4 Stereoselectivity origin

Stereoselectivity has been shown for Carvone and Limonene. In this section, we investigate the discrimination origin.

5.4.1 The vial differentiation

In the experimental protocol in Table 5.1, two vials are booked for the same VOC, namely Butanol. The experiment is repeated twice over two long sessions. So, in total, we have samples coming from 4 different vials containing the same VOC. In Figure 5.11, we report the data of these 4 vials projected onto the two first Principal directions. Despite all our rigid controls on the empty setup (Section 5.3.1), the data is divided in 4 clusters, each cluster corresponding to a vial. This sensitivity is surprising and, to our knowledge, previous literature did not ever report such a bias. The origin of the vial differentiation is mysterious, but consequences are clear: is the observed differentiation between (R)-form and (S)-form, only related to a difference of vial?

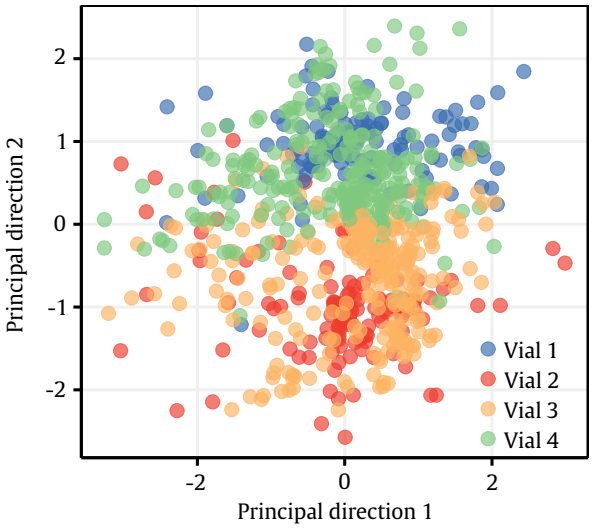


Fig. 5.11.: Different vials show different clusters. Each color stands for the data from a different vial containing Butanol.

First, inter-session analysis reduces this possibility. Indeed, we would be really lucky if the vial differentiation, random by nature, can generalize from Session 1 to Session 2 and vice-versa, and in addition, for both chiral molecules. We evaluate this probability to $\frac{1}{36}$. To do that, we simplify the problem by assuming that: we have drawn 4 vials independently from a common distribution and there exists a pair of vials which can generalize the knowledge learnt from this pair of vials to the other. The probability to randomly select this pair of vials is then: 1 out of $\frac{4\times3}{2} = 6$. So, if we would have considered only one chiral molecule in our study, the probability to obtain our results by luck would have been $\frac{1}{6}$. However, we demonstrated the same result for two different chiral molecules and these results are independent. The probability to obtain our two results by luck is therefore: $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

Second, we propose to go further in the analysis of the vial differentiation, session-by-session. We propose to look at the Euclidean distances between the samples of two different vials measured during cycle n. Since we have 6 different vials containing VOCs, this leads to 15 distances ($= \frac{6\times5}{2}$, the number of possible pairs) but only 6 among them are really relevant. For instance, looking both at the pair (Butanol vial 1, (R)-Limonene) and the pair (Butanol vial 1, (S)-Limonene) would be useless (only one of the two is enough). One of the 6 relevant pairs is the pair of (S)-Limonene and (R)-Limonene. Let us note x^R_{pn}(Limonene)

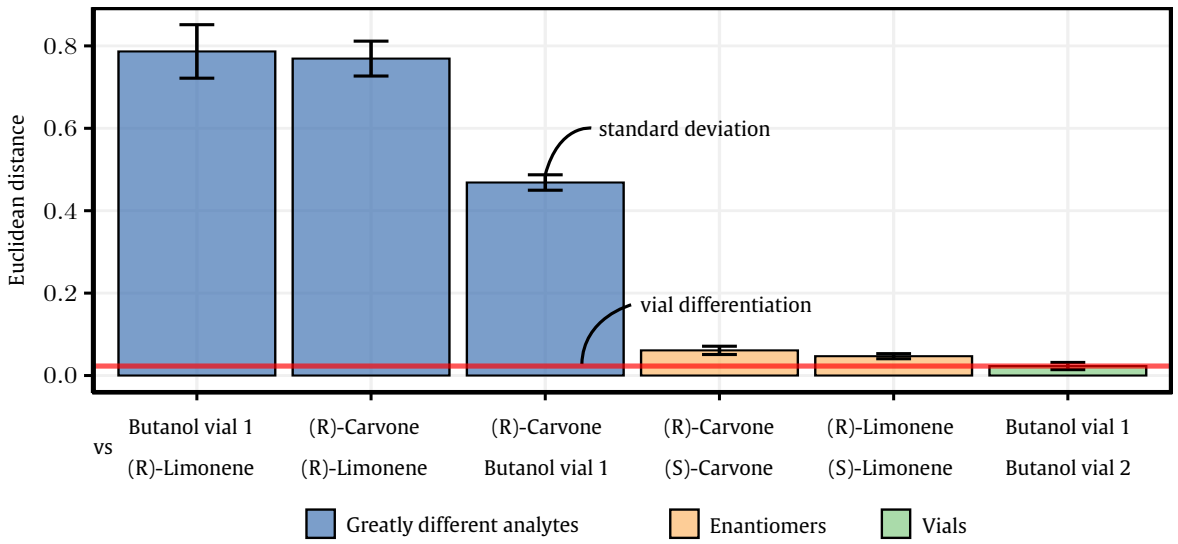


Fig. 5.12.: Vial differentiation analysis. We take as our criterion the Euclidean distance between samples from two different vials during the same cycle. These distances are computed for all the cycles of a single session. For each session, we compute the average and the standard deviation. Finally, the bars correspond to the weighted average over sessions and the error bars to the pooled standard deviation. The red line corresponds to the average distance between the two vials of Butanol (*i.e.* vial differentiation).

(resp. $x^S_{pn}(\text{Limonene})$) the normalized response of the chemical sensor p to (R)-Limonene (resp. (S)-Limonene) during cycle n. The Euclidean distance is defined as:

$$d_n(\text{(R)-Limonene}, \text{(S)-Limonene}) = \sqrt{\sum_{p=1}^{P} (x^R_{pn}(\text{Limonene}) - x^S_{pn}(\text{Limonene}))^2}$$

This distance is computed for each cycle n and separately for each session (since the two sessions do not have the same number of samples), without any drift correction, such as not to incorporate any bias. For each session, we average all the distances and compute the standard deviation. Finally, we average all the sessions by weighting with their number of samples. It corresponds to the bar ((S)-Limonene, (R)-Limonene) in Figure 5.12. The error bar corresponds to the pooled standard deviation. This process is repeated for each relevant pair of vials and results are reported in Figure 5.12.

Figure 5.12 confirms the vial differentiation since the distance separating the clusters of the same VOC (Butanol) in two different vials is not zero. In our case, pure vial differentiation is insufficient to explain the distance between enantiomeric samples.

We think that these checks should be considered by future studies, especially when the differences are tiny.

5.4.2 The chemical differentiation

Peptides are chiral molecules and are known to be good candidates for the differentiation of two mirror molecules [NS+17]. We sought to verify that it is indeed what drives stereoselectivity in our case, since our instrument also carries 2 achiral sensing materials. To do that,

we quantify the discriminative power of each sensing material. Recall that a given sensing material may be used for more than one chemical sensor, so that in the following, each sensing material is summarized by the average of its replicas. There are $M = 19$ different sensing materials on the device, with 17 of them chiral, and 2 achiral, the latter playing the role of controls. The goal of the analysis is to assess which sensing materials drive the discriminative power for the chiral compounds used here.

A simple analysis would look at the sensing materials one-by-one, and see if they can be used *on their own* for separating the two enantiomers. A problem arises rapidly: the normalization given in (5.1), compulsory for removing variations in measurements due to variations in concentration, implies the use of several sensing materials. This means we cannot look at sensing materials one-by-one. Another incorrect procedure would be to use all the sensing materials to first normalize and then to look at them one-by-one. This will implicitly introduce a bias invalidating the analysis.

We can illustrate it with a short numerical example using only two sensing materials for discriminating two VOCs, say A and B. We assume that sensing material 1 is discriminative whereas sensing material 2 is not. Note c_A and c_B the concentration of A and B during an experiment for which we have the raw responses reported in Table 5.3.

Sensing material 2 (with raw response $\tilde{x}_2$) is clearly not discriminative since it shares the same affinity with A and B, only a difference in concentration would lead to a difference in response. Sensing material 1 (with raw response $\tilde{x}_1$) is discriminative since its affinity with B is twice greater than with A. To get rid of the variations in concentration, we normalize using Eq. (5.1), leading to the normalized responses reported in Table 5.3.

	To A	To B
$\tilde{x}_1$	c_A	$2c_B$
$\tilde{x}_2$	c_A	c_B
x_1	0	$\frac{1}{2}\log(2)$
x_2	0	$-\frac{1}{2}\log(2)$

Tab. 5.3.: Numerical example. $\tilde{x}$ and x correspond to the unnormalized and normalized signature.

So, after normalization, sensing material 2 becomes discriminative, but only thanks to the initial discriminative power of sensing material 1. It is not hard to understand that this simple example generalizes to M sensing materials. In fact, only one discriminative sensing material can create discriminative information for all the others.

This bias is inevitable but we can reduce it by only looking at the sensing materials pair-by-pair. In the following, we assume that we have selected a pair of sensing materials, normalized the responses using (5.1) (for signatures of dimension 2) and corrected the drift. It is interesting to note that since in this case (5.1) creates a linear dependency between the two dimensions, we have effectively a one-dimensional measurement.

Second, we need to quantify the discriminative power of a given pair of sensing materials. We propose to use a kind of Signal-to-Noise Ratio (SNR). We define the signal part as the

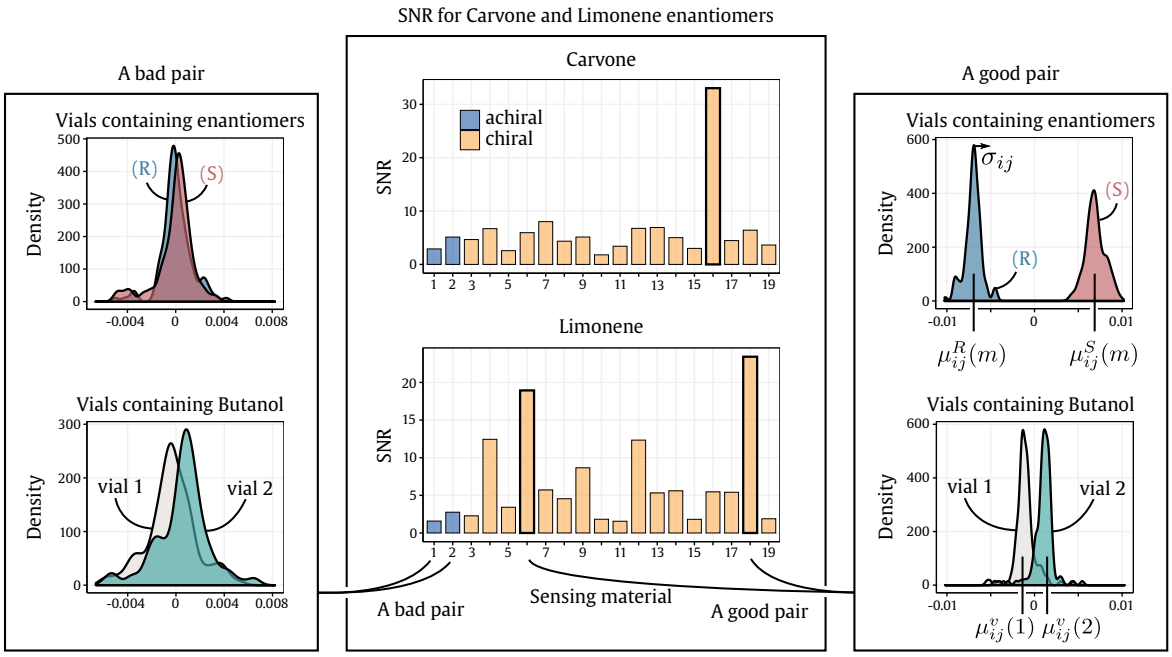


Fig. 5.13.: Signal-to-Noise Ratio (SNR) used to identify discriminative sensing materials. Middle figure, each bar corresponds to the averaged SNR of one sensing material across all its possible pairs. SNR value is averaged over the two sessions. The colors indicate either the sensing material is chiral (yellow) or achiral (blue). Left and right figures represent respectively the histograms (the data used for illustration is from Session 1) for a bad and a good pair of sensing materials.

squared distance between (R) centroid $\mu^R \in \mathbb{R}$ and (S) centroid $\mu^S \in \mathbb{R}$; the greater, the easier the differentiation between the 2 enantiomers. Noise can arguably be measured by the standard deviation σ of the samples (for a given VOC), which quantifies measurement uncertainty. σ is here calculated using samples from the (R)-enantiomer (the results do not change if we take the (S)-enantiomer). However, the previous section highlighted another kind of noise, namely vial differentiation, which is another source of experimental uncertainty. To incorporate this source of noise into the SNR, we focus again on the two vials containing Butanol (*cf* Table 5.1). Similarly to the signal estimation, we take as the noise attributed to vial differentiation, the squared distance between the centroid of the first vial $\mu^v(1) \in \mathbb{R}$ and the centroid of the second vial $\mu^v(2) \in \mathbb{R}$. The greater this value is, the more likely it is that discrimination performance is artifactual. Vial noise is then added to measurement noise. For a chiral molecule m and the pair formed by the sensing material i with the sensing material j, this gives:

$$\text{SNR}_{ij}(m) = \frac{(\mu_{ij}^R(m) - \mu_{ij}^S(m))^2}{(\mu_{ij}^v(1) - \mu_{ij}^v(2))^2 + \sigma_{ij}^2} \tag{5.2}$$

This estimation is computed for all 171 distinct pairs. To make the results more clear, we average the scores for a given sensing material over all pairs in which it appears. The resulting SNR and a graphical explanation are given in Figure 5.13. The greater the SNR, the more discriminative the sensing material. Due to confidentiality concerns, the nature of the sensing materials is not given and labels are substituted by arbitrary numbers.

We warn that this method does not really correspond to the “true" discriminative power of a sensing material since we are using pairs. However, we claim that this method is sufficient for evaluating which sensing materials are discriminative and which are not.

Results are represented in Figure 5.13 and show very high scores for sensing material 16 for Carvone and the two sensing materials 6 and 18 for Limonene. First, the 3 sensing materials are peptides and we observe that the two achiral molecules are always among the worst sensing materials. So, the controls did their job. Second, results highlight what we mentioned in Introduction of this chapter: chirality of the sensors is a necessary condition but not sufficient for stereoselectivity [Ben06]. Indeed, stereoselectivity is explained only by a couple of peptides and most of the peptides are not discriminative for the two chiral molecules studied. Finally, the 3 sensing materials could be replicated multiple times in future, to design a specific instrument to a given chiral pair.

5.5 Conclusion and future works

Stereoselectivity is the ability of a system to differentiate between two enantiomers. Due to the chiral character of Nature, stereoselectivity is a property of many biological systems. However, it remains the hardest problem in molecular recognition for man’s machines.

In this chapter, we showed the stereoselectivity of Aryballe’s eNose for two chiral molecules (Carvone and Limonene). To do that, we proposed a robust and statistically reliable methodology based on a realistic gas sampling system and two different sessions over several days of measurement. It demonstrated the ability of a peptide-based optoelectronic nose to tell the 2 enantiomers apart.

During this demonstration, we highlighted some controls which should be done by other studies. We especially reported an experimental confound which is inevitable, named vial differentiation. We proposed a metric including this inherent differentiation to identify the most discriminative sensing materials.

What are the perspectives?

Our results show the stereoselectivity of the instrument regarding Carvone and Limonene. Of course, it does not mean that the instrument can differentiate any enantiomeric pair. Further studies should be conducted for demonstrating the stereoselectivity on a wider range of chiral pairs.

Further studies should also be conducted on enantiomeric mixtures. The synthesis of an enantiomer often starts from its racemic mixture and having an inexpensive and flexible device that can measure relative enantiomeric concentrations would be highly valuable for production.

Robot olfaction

6

Contents

The results of this chapter have been submitted to a journal paper:

[hal-02448737] P. Maho, C. Herrier, T. Livache, P. Comon, and S. Barthelmé
"*Real-time gas recognition and gas unmixing in robot applications*"
Submitted to: *Sensors and Actuators B: Chemical*
Under review n°2

"[...] the wind brought him something, a tiny hardly noticeable something, a crumb, an atom of scent; no, even less than that: it was more the premonition of a scent than the scent itself [...]; it continually eluded his perception, was masked by the powder-smoke of the petards, blocked by the exudations of the crowd, fragmented and crushed by the thousands of other city odours. [...] he moved away from the wall of the Pavillon de Flore, dived into the crowd, and made his way across the bridge. Every few strides he would stop and stand on tiptoe in order to take a sniff from above people's heads. [...] he smelled the scent, stronger than before, knew that he was on the right track, dived in again, burrowed through the throng [...]. Here, he stopped, gathering his forces, and smelled. He had it. He had hold of it tight. The odour came rolling down the rue de Seine like a ribbon, unmistakably clear, and yet as before very delicate and very fine."

Patrick Süskind, *Perfume, the Story of a Murderer*[1].

In this extract, Patrick Süskind describes the tracking of a scent by Grenouille. Grenouille acts like an animal looking for some food or like a predator hunting its prey. The sense of smell of Grenouille is unusual for a human, who is rather guided by sight than by scent. Olfaction is in fact a key sense used by most species of animals for locating an object in the environment.

Source localization with olfaction is not a trivial task. Despite his incredible sense of smell, Grenouille is about to lose the track of the scent several times. To follow the scent, Grenouille must take up two main challenges. A first fundamental challenge lies in odour recognition. The track begins and finishes with odour recognition which is essential for identifying the source of interest and for discriminating between different compounds. The extract illustrates perfectly the difficulty of odour recognition in environment. For instance, the odour can be present at different levels of concentration and can also be mixed with other odours, of interest or not. The extract also illustrates the second main challenge, which is the definition of a movement strategy to easily and quickly locate the source. The optimal strategy can be hard to find especially due to obstacles present in environment (*e.g.* the crowd) and to variations in the wind direction.

This animal behaviour has recently inspired a research field, called **robot olfaction**. To name a few applications, robot olfaction can be used for gas leak detection, demining or pollutant localization [IWM12]. In most cases, these applications are based on the recognition of given Volatile Organic Compounds (VOCs) present in environment. This recognition can be

[1] Translated from the German by John E. Woods.

achieved thanks to gas sensors specific to the targeted VOCs (*e.g.* [Mea+13]). However, L. Marques [MNA02] demonstrated that the use of an array of non-specific gas sensors, namely an electronic nose, can also be an interesting alternative approach.

Electronic noses (eNoses) have been often studied in controlled settings in which temperature, humidity and VOC concentration are kept constant. In contrast, an olfactory robot will continuously measure VOCs with time-varying concentrations guided by diffusion and advection, and that under unstable environmental conditions [Tri10]. In a realistic environment, several different gas sources can be present, only one of which is of interest. This makes the task of **classification** fundamental. In addition, gases can mix in different proportion. Finding the identity of the mixing gases and their concentration is a quite complicated task which is referred to as **gas unmixing** or **Source Separation**.

In the literature, few works have been devoted to VOC classification with multiple different VOCs in uncontrolled environments [MT11] and most of them restrict their studies to 2 gas sources. Even in this case, one major assumption is that gas sources are sufficiently far away from each other to avoid interaction, meaning that there is no gas mixture (*e.g.* [Fan+19]). This assumption is quite unrealistic in real-life applications.

In this chapter, we tackle two issues. First, we deal with classification in uncontrolled environment where multiple gas sources of different VOCs are disseminated. Second, we study different robotic scenarios in which gas mixtures occur and we propose an unmixing algorithm.

In Section 1.3.3 (Chapter 1), we reviewed several key results obtained by the CREAB team [Bre18; Hur+19; Wee+19; Wee+20]. Their experiments were carried out with a lab instrument under lab conditions. In the same vein, in Chapter 5, we presented results under controlled conditions to show the stereoselectivity of Aryballe's eNose. By devoting this chapter to robot olfaction, we especially aim at studying this eNose in a complex environment in which environmental conditions and gas sampling can no longer be reliably controlled.

To that end, we have built two different robotic platforms. This kind of experimental setups is sometimes called **open sampling systems**, since the eNose is directly exposed to the environment to be monitored [Tri+11]. The two open sampling systems are introduced in Chapter 2 and are placed in an indoor environment with low advection (*e.g.* no ventilation system). One setup is used for assessing the recognition of 12 different VOCs disseminated in 24 isolated gas sources. The other setup is used for testing the unmixing of binary and ternary gas mixtures. These mixtures are generated due to either the proximity of two isolated gas sources or to the succession of scented trails. We proposed a real-time processing pipeline, able to deal both with gas recognition and with gas unmixing. The approach is based on a linear model with sparsity and smoothing constraints. It assumes the existence of a known dictionary.

The chapter starts with a short review of the main studies which deal with VOC classification and gas unmixing in the robot olfaction field. Then, the experimental section presents the different scenarios studied and several data sets generated during the thesis. The real-time processing pipeline is presented afterwards as well as the classification criterion used for quantifying the results. Finally, we present in two separated section the results obtained by the proposed method for dealing with the classification of 12 VOCs and with gas unmixing.

6.1 State of the art

In this section, we review the main studies related to the classification of multiple gas sources (*i.e.* at least 2 different VOCs) present at the same time in an open environment. To lighten the review, any study that only considers a single pure gas source or any study with multiple VOCs but presented in separated experiments is not detailed here. As electronic noses are likely to be weakly-specific devices, the use of highly-specific instruments with low cross-sensitivities to the target VOCs is out of the scope of this chapter (see for example the urban air quality study of [Mea+13]).

The literature on robot olfaction describes different algorithms, assessed with different setups which are not comparable amongst themselves [MT11]. In this section, we decided to split VOC classification in open sampling systems into two main types of work: first, the works assuming that the gas sources are far from each other (we define such gas sources as **isolated gas sources**) and second those dealing with mixtures. We already emphasize that the last one is an application much more rarely considered in the literature. Although it is the most realistic, it is also the most challenging.

6.1.1 Isolated gas sources

One of the earliest works is probably the one of Loutfi et al. [Lou+05; LC06]. They studied the classification of VOCs, up to 5, which were placed in cups present in the environment. With the help of visual clues, their approach was based on an approach similar to a classical 3-phase sampling. They first waited 30 seconds for the baseline acquisition and then injected the VOC during 20 seconds by placing the robot near the gas source. Finally, they waited for the total recovery of the chemical sensors before carrying another measurement. This approach implies that one waits for chemical sensors to recover, which can easily take 2-5 mins for MOS-based systems.

Then, much effort has been dedicated to use only transient responses, meaning only the measurement $\boldsymbol{y}(t)$, but at the cost of a limited number of gas sources. Loutfi et al. [Lou+09] placed two gas sources in two long corridors and improved a kernel-based approach, first proposed by Lilienthal and Duckett [LD04], for mapping the gas concentration of each VOC. This algorithm has been further improved by several studies from the same group [HB+14; Fan+18; Fan+19]. Monroy et al. [Mon+16] examined the use of time windows instead of single time points. They concluded that algorithms could be improved by taking advantage of the temporal correlation of samples in uncontrolled environment. In the same vein, F.

M. Schleif et al. [Sch+16] proposed an algorithm to deal with short time sequences, called generative topographic mapping through time [BHS97]. They validated their approach with a robotic arm and four different chemical substances which were presented sequentially to a MOS-based system.

Another line of research, in the absence of gas mixtures, is the examination of the effects of external parameters on classification accuracy. Palacìn et al. [Pal+19] studied the case where heating, ventilation and air conditioning were activated or not. They concluded that these parameters can help to locate gas leaks in open environments. Monroy et al. [MGJ17] examined the influence of robot speed for the discrimination of 2 gas sources. By using a MOS-based system, they particularly emphasized a loss in classification accuracy, up to 30%, when the motion speed strongly differs between the training samples and the test samples. Vergara et al. [Ver+13] made similar conclusions but regarding wind effects. They considered the recognition of 10 VOCs with different wind speeds under tightly-controlled conditions (pressure, humidity, temperature, concentration). An interesting aspect of their study is the number of VOCs used which is uncommon in the literature when one deals with open sampling system. However, in the case of this study, each VOC was measured in separated experiments.

An interesting conclusion of this short review is that most of the works consider only the case of a very limited number of different gas sources, which can be restrictive in practice.

6.1.2 Gas mixtures

The paper of Hernandez Bennetts et al. [HB+14] is probably one of the most realistic applications in the field. They considered 2 gas sources, separated by a small distance. They proposed an improved version of an existing kernel-based algorithm [LD04] which can deal with mixtures with the help of a PhotoIonization Detector (PID, see Section 1.1 for an explanation about the working principle of a PID). However, few experiments were carried out and one of them concluded to the low probability of one VOC just near its source location, which is counterintuitive. This could be explained by the exhaustion of the gas source. The authors emphasized the difficulty of obtaining a ground-truth in such scenarios.

Most existing works concerning gas unmixing in uncontrolled environment rely on **mixtures-learning**. Mixtures-learning means that authors generate all possible mixtures with the VOCs of interest, measure the instrument's response, and then learn (*e.g.* with a neural network) from the measured responses to predict possible responses on novel mixtures. Marques et al. [MNA02] built a setup in which an Ethanol source was disturbed by a Methanol source in a turbulent regime. Their solution was then based on the training of a neural network for locating the Ethanol source. Fonollosa et al. [Fon+14] also studied turbulent gas mixtures and proposed to use an Inhibitory Support Vector Machine [Hue+12]. A MOS-based system was placed in a wind tunnel and the goal was then to identify the presence of Ethylene when an interfering volatile was present. In another work, Fonollosa et al. [Fon+15] studied the composition of binary mixtures with an artificial neural network called Reservoir

Computing [MNM02]. The sensors were placed in a 60 mL measurement chamber and they continuously injected random mixtures of two compounds (including "pure" mixtures) with random transition times. The approach based on Reservoir Computing performed a better estimation of the true concentration of each compound compared to simpler methods, namely a linear regression and a Support Vector Regression.

An implicit drawback of a mixtures-learning approach for gas unmixing, is that it requires the generation of a lot of different mixtures for training the algorithm, ideally all the possible combinations of the studied VOCs. It can be manageable when one studies only binary mixtures but it rapidly becomes impossible when one deals with more complex mixtures. For instance, let us assume that we have R VOCs in the environment which can all mix in different proportions (including a null concentration). Then we have to generate $\Omega = 2^R - 1$ mixtures[2] to be sure that we can recognize any mixture of these R VOCs. For $R = 2$ as in the article of Fonollosa et al. [Fon+15], we have to manage "only" 3 different kinds of mixture (2 pure, 1 real mixture) at different concentration levels. But for $R = 3$, 7 different kinds of mixture have to be generated which can be more challenging regarding the experimental practicability. If we continue, for $R = 4$, we have to manage $\Omega = 15$ kinds of mixtures, for $R = 5$, $\Omega = 31$, etc...

We believe that a more valuable algorithm, easier to use in practice, would deal only with experiments generated from the pure VOCs. From these experiments with pure VOCs, we would build a dictionary and would try to unmix any mixture from this dictionary. In this way, only R experiments have to be generated which considerably reduces the experimental time (and human factors at the same time).

As a conclusion, little work has been done concerning the use of an electronic nose for dealing with mixtures in an uncontrolled environment.

6.2 Experimental

6.2.1 Optolectronic nose

Optoelectronic nose. Aryballe's eNose is detailed in Section 1.3. The instrument used in this chapter boards 19 different sensing materials which are replicated 3 or 4 times on the surface, leading to an array of $P = 59$ chemical sensors.

Main advantages for a robotic application. Most existing works in robot olfaction use MOS-based systems [IWM12]. So, we can highlight some advantages of using Aryballe's eNose for a robotic application compared to MOS-based sensors.

First, the number of chemical sensors is large compared to other robot studies. Using MOS-based sensors, this number generally reaches a maximum 5 or 6 elements, replicated or not. Here, we have four times more sensors since we have 19 different molecules on the sensing

[2] Each VOC is either present in the mixture or absent. There are therefore 2 possible states for each VOC, so 2^R distinct combinations including the "null" mixture.

surface (replicated 3 or 4 times). It is well known that, when one deals with gas mixtures, the number of sensors is a crucial parameter when one wants to identify the components of mixtures [CJ10] (see for instance Section 3.3.4 in Chapter 3 which presents a study about the influence of the number of sensors in a Blind Source Separation framework).

Second, this instrument demands much less power than a MOS-based system. For instance, let us consider a MOS gas sensor such as the Figaro TGS 2600 which is often used in the literature for detecting Carbon monoxide or Hydrogen. If only the power consumption for the heater part (which is the most power demanding by far) is considered, then according to the product's sheet, power consumption reaches 210mW [Fig13]. By doing a simple multiplication and assuming that we want 60 replicas of this sensor for our robot, we reach 12.6 W. By comparison with our instrument for which the working principle only needs a camera and an LED, we obtain 1.56 W (these 2 numbers do not take into account the on-board electronic system such as the pump which creates the flow of air). Thus, this instrument reduces the power consumption by a factor $\sim$8 which is absolutely not negligible in a mobile robot application for which battery life is crucial. In addition, with this instrument, we can easily add new sensors to the prism surface without increasing system complexity [Bre+18], meaning that we can double the number of sensors without changing power consumption.

Finally, a common concern with eNoses is stability over the medium and long term [MGG12]: are signatures stable enough that VOCs can be reliably recognize over weeks or months? In Chapter 4, we deeply reviewed this issue, called drift issue. Using a robotic setup (similar to the one in the next section), we ran repeated measurements of three different VOCs over several months and under different environmental conditions. The signatures did indeed drift over time, and we compared several correction methods. The results show that the signatures used for classification can be used without re-training over a period of several months. Other advantages have been highlighted in Section 1.3.3 (Chapter 1).

All these characteristics match the desirable attributes needed for a robot application described by R. A. Russell [Rus01].

6.2.2 Isolated gas sources

The first open sampling system is presented in Chapter 2, in Section 2.3.1. We called this setup Sniffer robot. The platform (see Figure 6.1) consists of a robot that carries the previously introduced eNose, moving over a flat surface where gas sources are placed. To have a better idea of the setup and of the experiments, we encourage the reader to watch a video[3] that I have made during the thesis and available at:

`http://simonbarthelme.eu/personal-website/videos/robot12VOC.mp4`

A funnel-shaped support has been made with a 3D printer and the PEEK (PolyEtherEtherKetone) injection tube of the eNose is inserted in this support in order to increase the suction

[3] If the link does not work, try this one:
`https://drive.google.com/file/d/1AddmCYFwQcHtG1aTksn8zyC-0e7CDzbU/view`

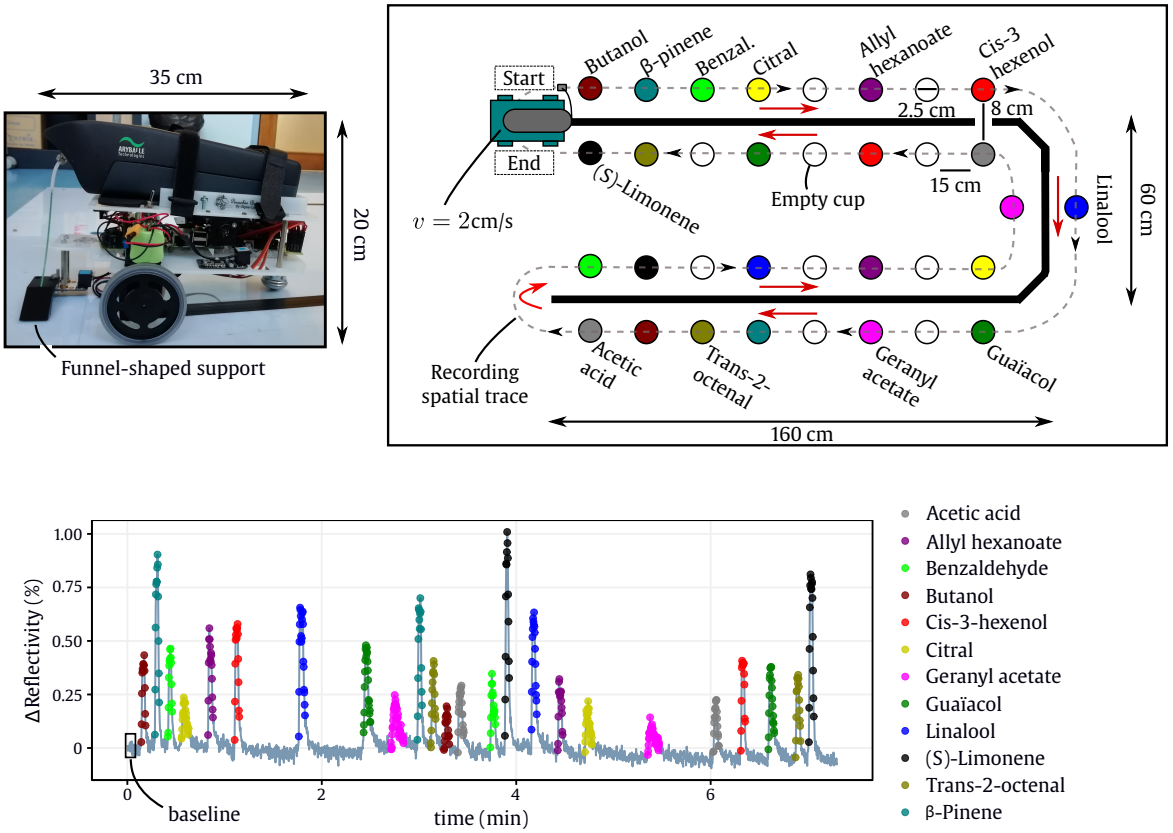


Fig. 6.1.: Sniffer robot platform. Top Left figure: the robot and some dimensions. Top Right figure: the experimental platform and some dimensions. 12 VOCs are disseminated in 24 gas sources along a path followed by the robot. Bottom figure: a time series of one chemical sensor during one lap (with baseline subtracted, taking the first points). Each peak corresponds to the passage of the robot over a gas source (a small cup). To extract the peaks, we perform a segmentation step detailed in Appendix A. The results of the segmentation is shown by the colored points (each color stands for one of the 12 VOCs).

area. The ground is a 2m $\times$ 1m $\times$ 2.5mm polycrystal plate which is lifted by 1.5cm. On this plate, a black path is drawn for the robot to follow. Along this path, the plate is pierced at 34 different locations thus enabling to slide below the plate up to 34 gas sources at the same time. Here, gas sources are small cups in which liquid solutions of VOCs are placed. This setup is placed in an indoor environment, basically a normal office with low natural advection (no ventilation system). The liquid solutions ($\sim$250μL) are put in the cups just before the experiment. The speed of the robot is set to 2 cm/s, the frame rate of the camera to 5 Hz and the airflow to 60 mL/min. The frame rate can be increased and the airflow can be decreased but these values are sufficient in practice to measure the chemical reactions and their dynamics.

The 12 VOCs we selected are listed in table 6.1. This selection has been only based on product availability and on safety, but *not* on whether their signatures are easy to tell apart. Each VOC is repeated twice along the track. The position of each VOC is optimized manually according to two criteria: first, we have to limit the desorption of one compound on the next compound and second, two cups of a VOC must have different neighbors. In this way, each gas source is

VOC	Formula	Molar mass (g/mol)	P_{sat} (mbar)	Chemical family
(S)-Limonene	$C_{10}H_{16}$	136.2	2.1	Alcene
β-pinene	$C_{10}H_{16}$	136.2	3.2	Alcene
Allyl-hexanoate	$C_9H_{16}O_2$	156.2	0.91	Ester
Geranyl acetate	$C_{12}H_{20}O_2$	196.3	0.035	Ester
Butanol	$C_4H_{10}O$	74.1	8.9	Alcohol
Cis-3-hexenol	$C_6H_{12}O$	100.2	1.4	Alcohol
Linalool	C_7H_6O	154.2	0.021	Alcohol
Benzaldehyde	C_7H_6O	106.1	1.7	Aldehyde
Trans-2-octenal	$C_8H_{14}O$	126.2	0.73	Aldehyde
Citral	$C_{10}H_{16}O$	152.2	0.12	Aldehyde
Acetic acid	$C_2H_4O_2$	60.0	21	Acid
Guaïacol	$C_7H_8O_2$	124.1	0.24	Phenol

Tab. 6.1.: List of the VOCs used and some of their properties (the vapor pressure P_{sat} was estimated at 25°C).

sufficiently distant from its neighbors which limits the mixing but does not completely prevent it. This also explains why some cups are left empty in Figure 6.1 and why "*only*" 24 cups out of 34 are really occupied by a VOC. The path is repeated 28 times (the total duration of the experiment is then 3h and 44 min), so for each VOC we have at most 56 peaks (such the ones represented in Figure 6.1).

Finally, a raw time series (with baseline subtracted, first points are taken for baseline estimation) of one chemical sensor during one lap is represented in Figure 6.1. Each peak corresponds to the passage of the robot over a gas source (so we can count 24 peaks).

The robot is not capable of interacting with the eNose yet. Ideally, we should know at least the location of the robot over time to relate the response of the eNose to a spatial position. However, we do not have this information neither. So, we do not know exactly when the robot is over a gas source, except that it must correspond to a peak measured by the instrument. To extract the peaks, we perform a segmentation step detailed in Appendix A. The results of the segmentation is shown by the colored points (each color stands for a VOC) in Figure 6.1.

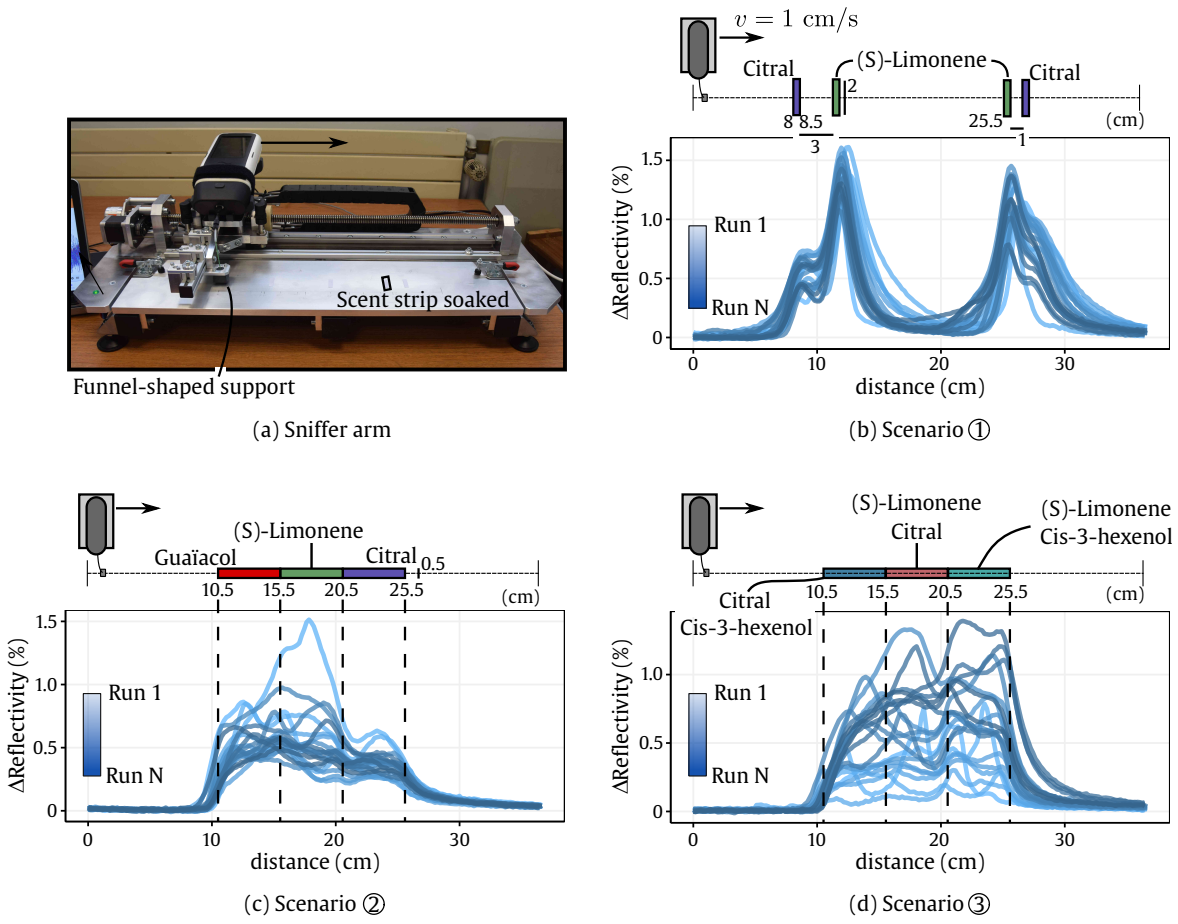
6.2.3 Gas mixtures

The second open sampling system is presented in Chapter 2, in Section 2.3.2. We called this setup Sniffer arm. This setup is used to evaluate different realistic scenarios that can be encountered in a robot olfaction application.

The main part is an aluminium trapezoidal shaft that moves the eNose along one dimension. To have a better idea of the setup and of the experiments, we encourage the reader to watch a video[4] that I have made during the thesis and available at:

http://simonbarthelme.eu/personal-website/videos/closeSources.mp4

[4]If the link does not work, try this one:
https://drive.google.com/file/d/14exr-R1c4hTyitn1WeuqKYA3uLgDrI5z/view



(a) Sniffer arm (b) Scenario ① (c) Scenario ② (d) Scenario ③

Fig. 6.2.: Sniffer arm setup. (a) The setup. (b) Scenario ①: isolated gas sources are spatially close. (c) Scenario ②: successive trails of gas sources (pure VOCs). (d) Scenario ③: successive trails of gas sources (binary VOC mixtures). Below each scenario, we represent the averaged ΔReflectivity (across chemical sensors) for each sweep (first sweep is the lighter). More details are given in the text.

The total length of this path is 36.5cm. With this setup, we can perform sweeps multiple times, from left to right and vice versa. To save space, we detail only the data from one direction of sweep, but the rest is presented in Appendix C.2 (the two directions are highly similar). Again, a funnel-shaped support is added to increase suction area. Gas sources are now scent strips, directly placed along the 1D-path, on which few drops are deposited. The speed of arm movement is set to 1 cm/s.

In the literature on gas recognition in an uncontrolled environment, authors generally work only with pure compounds which are far from each other in order to avoid mixing. We believe that these restrictions are not realistic in the field. That is why we focus in this study on binary, or ternary, mixtures. To this end, we propose different realistic scenarios with increasing complexity:

- **Scenario** ①, Figure 6.2b: gas sources are spatially close which leads to gas mixtures. We consider two scent strips which are separated either by 3 cm or by 1 cm. The two gas sources are Citral and (S)-Limonene. One drop of their liquid solutions (∼50 μL) is deposited on the scent strips. The averaged ΔReflectivities of 20 sweeps on the Figure

6.2b clearly indicates that the 2 gas sources mix. Unsurprisingly, the amount of mixing depends on the distance separating the two gas sources.

- **Scenario** ②, Figure 6.2c: gas sources are trails which are successive. We use three scent strips (5 cm long) which are placed one after the other. The first trail contains Citral, the second contains (S)-Limonene and the last one contains Guaïacol. Three drops of their liquid solutions ($\sim$150 μL) are deposited on the scent strips. The averaged ΔReflectivities on the Figure 6.2c of 20 sweeps clearly demonstrate the complexity of the task in which transitions between VOCs are unclear.
- **Scenario** ③, Figure 6.2d: gas sources are again trails but no longer pure compounds. We consider three scent strips (5 cm long) which are placed one after the other. The first trail contains Citral and Cis-3-hexenol, the second contains Citral and (S)-Limonene and the last one contains (S)-Limonene and Cis-3-hexenol. For a given scent strip, three drops of each liquid solution ($\sim$150 μL) of the 2 compounds are deposited on the scent strip. Thus, the mixtures operate both in the gas phase and in the liquid phase. Again, the averaged ΔReflectivities on the Figure 6.2d of 20 sweeps clearly demonstrate the complexity of the task which requires unmixing algorithms that can deal with ternary mixtures.

In each one of these scenarios, the experiments involve placing the scent strips, depositing the compounds and doing several sweeps (N=20). At the end of each sweep, the arm stays in place for 20 sec. The response of the eNose during the sweeps is recorded continuously (without any interruption between each sweep). Compared to the previous setup, we know when the arm reaches the end of a sweep. The segmentation of the signals is therefore straightforward and does not require any additional post-processing. We extract for each sweep only the part of the signal during which the arm is moving.

The spatial scale of this setup is limited compared to other experimental setups that can be found in the literature. However, in large-scale setups it is very difficult to obtain a ground truth [HB+14]. Indeed, nobody can tell if the results of the proposed algorithm far from the gas sources are valid or not since no one can tell exactly the proportion of a VOC at these distances in a realistic environment. Thus, we consider that a small-scale setup is completely appropriate since we would not have more information with a larger one.

6.3 Real-time unmixing

We propose a single processing pipeline which we use in all the scenarios previously introduced. It can be used for the recognition of isolated gas sources (as for Sniffer robot, described in Section 6.2.2) and for the unmixing of binary and ternary mixtures (as for Sniffer arm, see Section 6.2.3).

Requirements for real-time algorithms. Recently, [Fan+19] highlighted the need for real-time algorithms in a robot olfaction application such as emergency response scenarios.

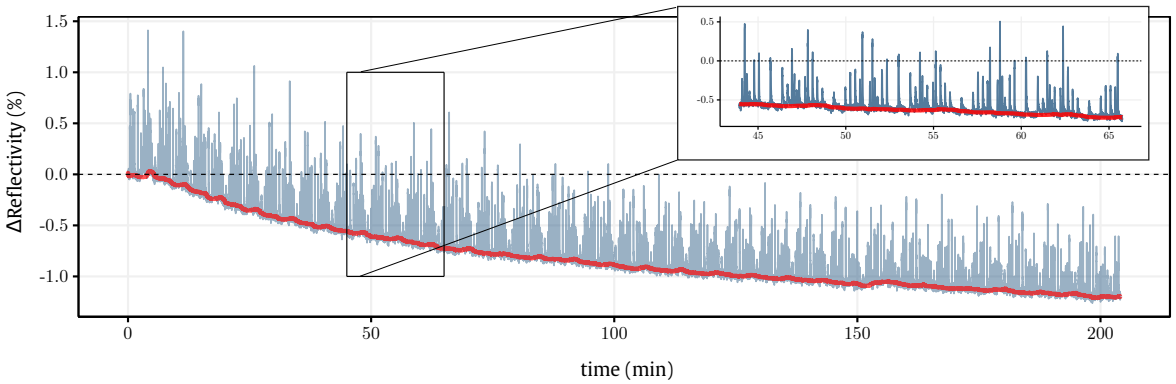


Fig. 6.3.: Baseline drifts over time. Drift estimation is represented in red.

Real-time algorithms must be of low algorithmic complexity to avoid any processing lag, and the processing pipeline described here is designed with that constraint in mind.

Beyond the constraint of low algorithmic complexity, real-time data-processing algorithms must be causal: at the current time point t, we can only use the data acquired up to time t. All the results shown include this constraint[5]. Another direct implication of the real-time constraint is that the whole processing pipeline for the current time point t must have a computation time lower than the sampling period (here, 200 ms). We report computation time below, but need to point out that it depends on the data, on the programming language (here, `R`) and on computer performance (Dell Inc. Latitude XT3, CPU i7-2640M 2.8GHz, Ubuntu 16.04). The timings we report could be greatly reduced, as `R` in particular is an interpreted language that we use for convenience. An order-of-magnitude improvement is to be expected from implementing the algorithms in a compiled language.

Baseline drift. In Chapter 2, in Section 2.3.3, we discussed an issue encountered with the robotic setups: the baseline is drifting over time. To be more precise, baseline suffers always from drift whatever the setup. Using a 3-phase sampling, the baseline drift is easily corrected since some time is saved for baseline acquisition. However, with our robotic setups, the eNose is continuously exposed to the environment and there is no specific time saved for acquiring the baseline. Consequently, the baseline drift appears at the same time as the VOC injection. To solve the problem, in Section 2.3.3, we use a q-quantile filter on a time window of size k which estimates the baseline trend (see Figure 6.3). This estimation is then simply subtracted. This filter can be applied in real-time since the computation time required is around 10 ms by setting the parameters to: $k = 100$ seconds and $q = 0.1$. The parameters have been manually selected for the experiments. Automatic tuning of these parameters is an interesting topic for further research. In the following, we assume that the baseline drift of each chemical sensor p at time t has been estimated and subtracted from the response.

Notation. To set notation, we note $y_p(t)$ the time series of the chemical sensor p. As in previous chapters, we note respectively N_t, P and R, the duration of the recording, the number

[5]Performance could be improved by processing the data in a batch, but then the algorithms would not be real-time anymore

of chemical sensors and the number of VOCs. As explained in the introduction, the instrument generates a signature $\boldsymbol{k}_r \in \mathbb{R}^P$ for each VOC r (the extraction process is detailed in this chapter, in Section 6.5.1). We stack all these signatures into a matrix $\boldsymbol{K} = [\boldsymbol{k}_1, ..., \boldsymbol{k}_R] \in \mathbb{R}^{P \times R}$. The intensities $\boldsymbol{c}(t) \in \mathbb{R}^R$ at time t are stacked into a matrix $\boldsymbol{C} = [\boldsymbol{c}(1), ..., \boldsymbol{c}(N_t)] \in \mathbb{R}^{R \times N_t}$ and measurements are stacked into a matrix $\boldsymbol{Y} = [\boldsymbol{y}(1), ..., \boldsymbol{y}(N_t)] \in \mathbb{R}^{P \times N_t}$.

6.3.1 Model

As described in Section 6.1.2, a solution proposed by some authors is based on mixtures-learning. We already discussed the main drawback of this strategy. In this chapter, we propose to use a known dictionary $\boldsymbol{K} \in \mathbb{R}^{P \times R}$ containing all the signatures of the R studied VOCs. This method requires little experimental time (in the training phase), as only pure VOCs have to be measured. The goal is then to estimate a vector $\boldsymbol{c}(t) \in \mathbb{R}^R$ which gives the intensity of each VOC at time t. This implies the formulation of a model relating $\boldsymbol{K}$ and $\boldsymbol{c}(t)$ to the measurement $\boldsymbol{y}(t)$.

The model we formulate is **linear**, because linearity greatly simplifies the computations. We know from Chapter 3 that linearity cannot hold in general, but holds approximately in a low concentration or low affinity regime (see Section 3.3). In addition, there are several experimental pieces of evidence in previous chapters which support linear model. In Chapter 1, we reviewed several studies of authors working with the same type of eNoses (see Section 1.3.3). They showed that the response of the eNose is proportional to concentration [Bre18], at least for a range of concentration depending on the VOC. In Chapter 4, we used a linear model successfully to normalize signatures for concentration. Another experimental piece of evidence in favor of using linear models is provided in Section 6.5.1 in which we represent the linear fitting of data of a single VOC.

Given a vector of sensor responses at time t, $\boldsymbol{y}(t) \in \mathbb{R}^P$, we express the linear model both in the presence of a single VOC r and of a mixture of VOCs:

Single VOC	Mixture of VOCs	
$\boxed{\boldsymbol{y}(t) = \boldsymbol{k}_r c_r(t)}$	$\boxed{\boldsymbol{y}(t) = \boldsymbol{K}\boldsymbol{c}(t)}$	(6.1)

The model for a single VOC is used in Section 6.5.1 for the estimation of the signature $\boldsymbol{k}_r \in \mathbb{R}^P$. In the following, we focus on the linear mixing model which simply assumes that the measurement $\boldsymbol{y}(t)$ is a linear combination of the signatures of pure VOCs. Again, we emphasize that this method requires little experimental time since we only need to generate data with pure VOCs for the estimation of $\boldsymbol{K}$.

6.3.2 Optimization problem

Despite its simplicity, linear unmixing has been successfully applied in many fields such as remote sensing [BD09], fluorescence microscopy [Dic+01] or spectroscopy [Mou+06]. A simple way to estimate $\boldsymbol{c}(t)$ is to minimize a least squares cost function:

$$\hat{\boldsymbol{c}}(t) = \arg\min_{\boldsymbol{c}} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 \tag{6.2}$$

In this case, the solution is easily obtained from the pseudo-inverse of $\boldsymbol{K}$. However, the latter analytic solution may produce negative intensity values, which is not physically possible. To avoid this, we add a **non-negativity constraint** on $\boldsymbol{c}(t)$. In addition, we do not expect that all the VOCs from the dictionary are present in the mixture (recall that $R = 12$ and that we study ternary mixtures in the worst case). This implies that we need to add a **sparsity constraint** on $\boldsymbol{c}(t)$. Sparsity can help the estimation since $\boldsymbol{K}$ can suffer from an ill-conditioning due to correlated signatures. The sparsity of $\boldsymbol{c}(t)$ can be naturally imposed by the ℓ_0-"norm" which counts the number of non-null elements in $\boldsymbol{c}(t)$:

$$\hat{\boldsymbol{c}}(t) = \arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 \text{ s.t. } \|\boldsymbol{c}\|_0 \leqslant \gamma \tag{6.3}$$

Here, it implies that at most γ intensity values of $\boldsymbol{c}(t)$ will be non-null (we call **support** $\mathcal{S}(t)$ of $\boldsymbol{c}(t)$ the position of non-null coefficients). However, this optimization problem is known to be NP-hard and the global solution can be reached only in small-scale applications. Indeed, worst-case computation time increases exponentially with the size R of the dictionary. Another approach, which we choose here, is to relax this constraint with the ℓ_1-norm, $\|\boldsymbol{c}\|_1 = \sum_{r=1}^{R} |c_r|$, through a penalty term weighted by $\lambda \geqslant 0$:

$$\hat{\boldsymbol{c}}(t) = \arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1 \tag{6.4}$$

Finally, we add a last constraint to improve the solution, which we call **support continuity**. Given a time t and a small integer ω, the supports $\{\mathcal{S}(t-\omega), ..., \mathcal{S}(t)\}$ of $\{\boldsymbol{c}(t-\omega), ..., \boldsymbol{c}(t)\}$ must be similar. This constraint implies a weak assumption which is that during a small time frame (defined by ω) the composition of the mixture does not change much (here, $\omega = 1$ sec). This constraint could be formulated as an additional penalty term such as $\mu\|\boldsymbol{c} - \boldsymbol{c}(t-1)\|_2^2$ which would smooth the estimation of $\boldsymbol{c}(t)$ in relation to $\boldsymbol{c}(t-1)$. However, this would require tuning an additional hyperparameter μ. Instead, we propose an heuristic that we detail in the next paragraph. For the rest, we define $\boldsymbol{c}^{\mathcal{S}(t)}(t) \in \mathbb{R}^{\gamma}$ the restricted version of $\boldsymbol{c}(t)$ to the support $\mathcal{S}(t)$, $\boldsymbol{C}_{\omega}^{\mathcal{S}(t)}(t) = [\boldsymbol{c}^{\mathcal{S}(t)}(t-\omega), ..., \boldsymbol{c}^{\mathcal{S}(t)}(t)] \in \mathbb{R}^{\gamma\times(\omega+1)}$, $\boldsymbol{Y}_{\omega}(t) = [\boldsymbol{y}(t-\omega), ..., \boldsymbol{y}(t)] \in \mathbb{R}^{P\times(\omega+1)}$ and $\boldsymbol{K}^{\mathcal{S}(t)} \in \mathbb{R}^{P\times\gamma}$ the restricted version of $\boldsymbol{K}$ to the support $\mathcal{S}(t)$. We define the subproblem:

$$\hat{\boldsymbol{C}}_{\omega}^{\mathcal{S}(t)}(t) = \arg\min_{\boldsymbol{C}\geqslant 0} \|\boldsymbol{Y}_{\omega}(t) - \boldsymbol{K}^{\mathcal{S}(t)}\boldsymbol{C}\|_F^2 \tag{6.5}$$

where $\boldsymbol{C} \geqslant 0$ means that all the elements of $\boldsymbol{C}$ are non-negative and $\|\cdot\|_F$ stands for the Frobenius norm.

6.3.3 Optimization method

The proposed method is divided into two main blocks: first, estimate which VOCs have a non-null intensity in $\boldsymbol{c}(t)$, *i.e.* solve Problem (6.4), and, second, integrate the support continuity constraint.

Solve Problem (6.4). Let us assume that we have a measurement $\boldsymbol{y}(t)$ from which we want to identify $\boldsymbol{c}(t)$. The first problem is to find the support $\mathcal{S}(t)$ of $\boldsymbol{c}(t)$ (the position of non-null coefficients). This can be done by solving the optimization problem in eq. (6.4), in which the parameter λ influences the degree of sparsity of $\hat{\boldsymbol{c}}(t)$, meaning the estimated number of VOCs in the mixture $\boldsymbol{y}(t)$; the larger λ, the sparser $\hat{\boldsymbol{c}}(t)$. If λ is too large, then the algorithm may converge to the null solution ($\hat{\boldsymbol{c}}(t) = \mathbf{0}_R$). On the other hand, if $\lambda = 0$ then all coefficients of $\hat{\boldsymbol{c}}(t)$ can be non-null and only the data misfit will matter. λ is therefore a crucial parameter in the estimation of the support $\mathcal{S}(t)$ and there exists a λ, called λ_{min}, which correctly balances the tradeoff sparsity - data misfit. In order to estimate λ_{min} and solve the problem (6.4) for $\lambda = \lambda_{\text{min}}$, we use the algorithm proposed by [FHT10] and the associated `R` package `glmnet`.

The algorithm relies on a grid search and a β-fold cross-validation procedure. Given a λ, $\beta \in \mathbb{N}$ folds are generated from $\boldsymbol{y}(t)$, then β-1 folds are used for estimating $\hat{\boldsymbol{c}}(t)$ with a coordinate descent, and finally the remaining fold is used for prediction from which a mean squared error is estimated. This procedure is repeated for each fold, which gives an estimation of the mean squared error for the given λ. This procedure is repeated with the same folds for the sequence of λ. The sequence of values of λ is picked using the method proposed by [FHT10].

Cross-validation (CV) is a common technique to estimate the prediction performance of a model. In our case, it has the advantage of being sensitive to the number of non-null coefficients. Another approach would be to simply compute a mean squared error using the whole $\boldsymbol{y}(t)$ for fitting and predicting. However, with this approach, the less sparse $\hat{\boldsymbol{c}}(t)$ is, the better the fit is and the lower the mean squared error is. By using CV, the error is computed on data which have not been used for the optimization. So, more parameters (*i.e.* $\hat{\boldsymbol{c}}(t)$ is less sparse) in the model does not imply less error (in fact, these parameters will learn the noise in the training data which will negatively impact the performance on new data).

Finally, the value of λ corresponding to the minimum mean squared error is chosen as λ_{min}. With $\lambda = \lambda_{\text{min}}$, the algorithm then estimates $\hat{\boldsymbol{c}}(t)$ from the whole $\boldsymbol{y}(t)$ via a coordinate descent. The support $\mathcal{S}(t)$ is therefore identified from $\hat{\boldsymbol{c}}(t)$ as $\mathcal{S}(t) = \{r \in [\![1, R]\!] : \hat{c}_r(t) \neq 0\}$. We emphasize that this method has the great advantage of not requiring the knowledge of the number of VOCs in the mixture $\boldsymbol{y}(t)$ (this number is estimated through λ_{min}). In Algorithm 10, we do not describe again the cross-validation procedure for the estimation of λ in the minimization process (6.4) for lack of space.

Solve support continuity constraint. At this stage, we have an estimation of $\mathcal{S}(t)$ from $\hat{\boldsymbol{c}}(t)$ considering only the time t and we have already estimated the previous ω supports $\{\mathcal{S}(t-\omega), ..., \mathcal{S}(t-1)\}$. In order to integrate the support continuity constraint, we just find

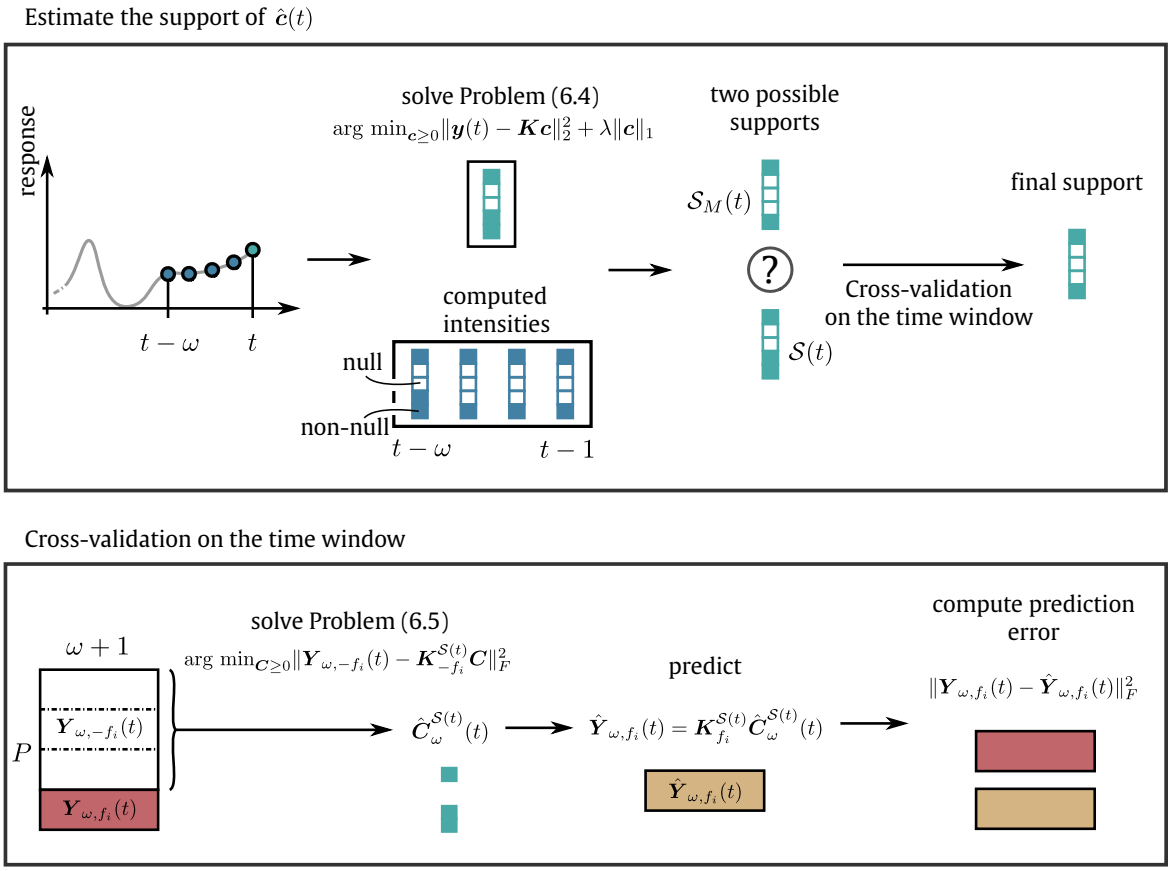


Fig. 6.4.: Graphical explanation of the estimation of the support. For estimating the support of the intensity vector $\boldsymbol{c}(t)$, we proceed in two steps. First, we solve Problem (6.4). Due to the sparsity constraint, $\hat{\boldsymbol{c}}(t)$ has many null coefficients. We call $\mathcal{S}(t)$ the position of non-null coefficients. Second, we compare $\mathcal{S}(t)$ with the previous supports estimated from the previous time points (in a small time frame). This step aims at including the support continuity constraint in the estimation of $\boldsymbol{c}(t)$. Support continuity means that we expect that the VOCs present in the mixture at time t may be the same than the VOCs present at time $t-\omega$ (for ω small). In other words, the constraint adjusts the support at time t by considering the previous estimated supports. For that, we compare $\mathcal{S}(t)$ with the most frequent support $\mathcal{S}_M(t)$. The comparison is done by cross-validation and the support having the lowest error on the time window is chosen.

the most frequent support $\mathcal{S}_M(t)$ of the $\omega+1$ supports. If the support $\mathcal{S}_M(t)$ equals $\mathcal{S}(t)$, then we go to the next step. Otherwise, we need to identify which support better explains the whole time window. To that end, we perfom again a β-fold cross-validation. We carry out the cross-validation on the rows of $\boldsymbol{Y}_\omega$ instead of its columns, since the two supports are known. We split the P rows of $\boldsymbol{Y}_\omega(t) \in \mathbb{R}^{P\times(\omega+1)}$ into β folds f_i, $i \in [\![1,\beta]\!]$ (f_i contains the indices of each fold). For each support and each fold, we solve the subproblem (6.5) considering $\boldsymbol{Y}_{\omega,-f_i}(t)$ (it contains the rows of $\boldsymbol{Y}_\omega(t)$ which are not in f_i), with a quasi-Newton algorithm called L-BFGS [Byr+95]. We then predict $\boldsymbol{Y}_{\omega,f_i}(t)$ (it contains the rows of $\boldsymbol{Y}_\omega(t)$ which are in f_i) and compute the sum of squared errors. We repeat the procedure for each fold, which gives an estimation of the sum of squared errors for each support. Finally, we identify $\mathcal{S}(t)$ as the support with the minimal error. For the first ω measurements, we just set ω to $t-1$. A graphical explanation of the estimation of the support is reported in Figure 6.4.

Algorithm 10 Real-time estimation of the intensities $\boldsymbol{c}(t)$

Require: ω, τ, β
  # Estimate the support of $\boldsymbol{c}(t)$
  $\hat{\boldsymbol{c}}(t) = \arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$
  $\mathcal{S}(t) = \{r \in [\![1, R]\!] : \hat{c}_r(t) \neq 0\}$
  # Homogenize the support in relation to the window
  Let $\mathcal{S}_M(t)$ be the most frequent support in $\{\mathcal{S}(t-\omega), ..., \mathcal{S}(t)\}$
  if $\mathcal{S}(t) \neq \mathcal{S}_M(t)$ **then**
    # Compute the cross-validation error for each support
    Split the P rows of $\boldsymbol{Y}_\omega(t)$ into β folds f_i, $i \in [\![1, \beta]\!]$. f_i contains the indices of the test fold i. $\boldsymbol{A}_{f_i}(t)$ contains the rows $\in f_i$ and $\boldsymbol{A}_{-f_i}(t)$ contains the rows $\notin f_i$.
    $J, J_M = 0$
    for $i = 1$ **to** β **do**
      $J = J + \min_{\boldsymbol{C}\geqslant 0} \|\boldsymbol{Y}_{\omega,-f_i}(t) - \boldsymbol{K}_{-f_i}^{\mathcal{S}(t)}\boldsymbol{C}\|_F^2$
      $J_M = J_M + \min_{\boldsymbol{C}\geqslant 0} \|\boldsymbol{Y}_{\omega,-f_i}(t) - \boldsymbol{K}_{-f_i}^{\mathcal{S}_M(t)}\boldsymbol{C}\|_F^2$
    end for
    if $J_M < J$ **then**
      $\mathcal{S}(t) = \mathcal{S}_M(t)$
    end if
  end if
  # Estimate $\boldsymbol{c}(t)$ *on the final support*
  $\hat{\boldsymbol{c}}(t) = 0_R$
  $\hat{\boldsymbol{c}}^{\mathcal{S}(t)}(t) = \arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}^{\mathcal{S}(t)}\boldsymbol{c}\|_2^2$
  # Smooth $\hat{\boldsymbol{c}}(t-\tau)$
  $\hat{\boldsymbol{c}}(t-\tau) = \frac{1}{2\tau+1}\sum_{i=-\tau}^{\tau} \hat{\boldsymbol{c}}(t-\tau-i)$

Eventually, the final step is then just to solve the problem (6.5) for the final support $\mathcal{S}(t)$, limiting the estimation to $\boldsymbol{y}(t)$. Furthermore, a smoothing filter is then applied to $\hat{\boldsymbol{c}}(t-\tau)$, where τ is an additional parameter (here, $\tau = 0.4$ sec). We simply replace $\hat{\boldsymbol{c}}(t-\tau)$ by the mean value of $\{\hat{\boldsymbol{c}}(t-2\tau), ..., \hat{\boldsymbol{c}}(t)\}$. In Appendix C.1, we provide an intensive comparison showing the interest of using a smoothing filter. We also show in Appendix C.1 the benefits of using all the constraints that we impose in our method.

The algorithm is summarized in Algorithm 10. It is important to note that this method does not assume that we know the number of VOCs in the mixture $\boldsymbol{y}(t)$ and does not normalize the data. So, the vector $\hat{\boldsymbol{c}}(t)$ which contains the intensity of each signature in the measurement $\boldsymbol{y}(t)$ should be related to the VOC concentration in some way. This claim must be confirmed by further experiments measuring the true VOC concentration with an additional instrument that can determine ground-truth concentration.

6.3.4 Computation time

In Section 6.5.4, we perform cross-validation in order to estimate a score for the data from Sniffer robot ($R = 12$ VOCs). Out of ~600 000 measurements $\boldsymbol{y}(t)$, the computation time of Algorithm 10 was then 75 $\pm$17 ms. Eventually, the computation time needed by the two algorithms is lower than the sampling period (200 ms), as required.

6.4 Evaluation

In this study, we have no way of measuring the *actual* concentration of each VOC, which limits the assessment of our method. In the following, we detail a classification-based criterion.

To compute a score, we can use the location of each gas source. In fact, time and distance are related in a trivial way. So, we can easily relate the time measurement $\boldsymbol{y}(t)$ and the estimated intensities $\hat{\boldsymbol{c}}(t)$ to a location along the path. Therefore, we can define a set of time points, say $\{t_1^r, ..., t_{N_r}^r\}$, for which the robot is over the gas source of the VOC r. Then, we can associate with each time t in this temporal range, a label $\ell(t)$ according to the maximum intensity:

$$\forall t \in \{t_1^r, ..., t_{N_r}^r\}, \quad \ell(t) = \arg\max_i \hat{c}_i(t) \tag{6.6}$$

From this label $\ell(t)$, we can then compute a score based on the position of the gas sources. Ideally, any label $\ell(t)$ matches with the identity of the gas source. However, this score can only work for pure compounds. Since we have also binary mixtures at a time t (Scenario ③ of Section 6.2.3), we extend the previous criterion to the mixtures of γ VOCs by simply taking the γ first maxima as the classification result:

$$\forall t \in \{t_1^{r_1,...,r_\gamma}, ..., t_{N_r}^{r_1,...,r_\gamma}\}, \forall j \in [\![1, \gamma]\!] \quad \ell_j(t) = \arg\max_{i \setminus \{\ell_1(t),...,\ell_{j-1}(t)\}} \hat{c}_i(t) \tag{6.7}$$

From the labels $\{\ell_1(t), ..., \ell_j(t)\}$, we consider the output of the algorithm a success if it correctly identifies all the VOCs present in the mixture. The output is considered false if at least one VOC has not been correctly identified.

We warn the reader that the location of the gas sources (diameter of the cups, length of the scent strips) does not really correspond to the beginning and to the end of interactions (meaning, the rise and the baseline return). Indeed, there exists some inherent lags, such as: a chemical lag due to the dynamics of the chemical reactions (adsorption, desorption), a lag due to the airflow (the delay in transporting one molecule from the floor to the prism surface), a lag due to diffusion (the spatial range is in fact greater than the spatial area of the gas source) and another lag due to the funnel. Consequently, the response can often spread over a spatial range larger than the location of the gas source. Future developments can integrate some deconvolution tools to compensate for all these lags (see for example the recent study of [MBM19]).

Evaluation for data from Sniffer robot (Section 6.2.2). In this setup, the spatial area of the cups and the corresponding temporal range are quite small. As a reminder, the cups have a diameter of 2.5 cm and the speed of the robot is 2 cm/s, leading to only 6 measurements (frame rate = 5Hz) if we consider strictly speaking the location of the gas source. To avoid an overoptimistic score (due to a too small temporal range), we extend the location of the gas sources and we take the time points corresponding to the entire peaks (for an illustration of the peaks, see Figure 6.1). For finding the peaks, we perform a segmentation of the signal.

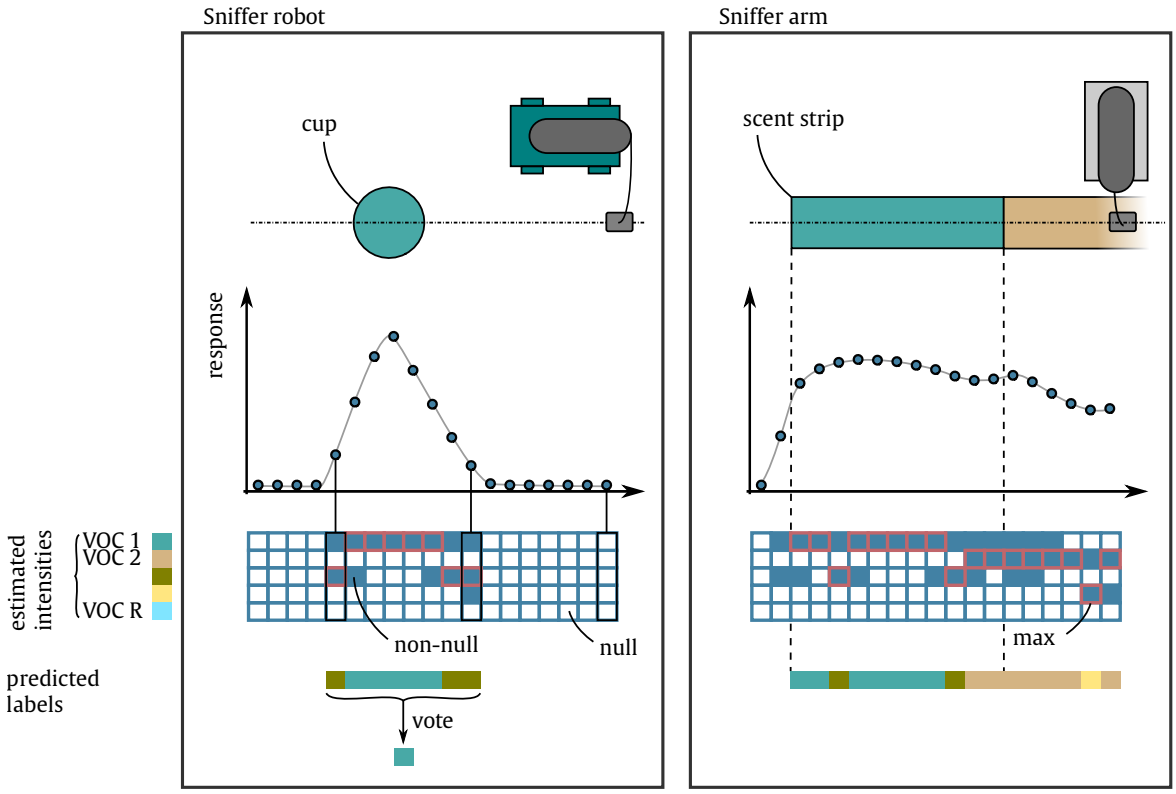


Fig. 6.5.: Graphical explanation of the evaluation procedure. We identify each point $\boldsymbol{y}(t)$ as being the VOC having the max intensity in $\hat{\boldsymbol{c}}(t)$. This label is then compared to the identity of the gas source over which the robot is passed. For Sniffer robot, we perform a vote between all the labels extracted and associate with the gas source the majority label. For Sniffer arm, we use each label $\ell(t)$ to compute the score.

This segmentation assumes that there is no mixture between two successive cups (which may not be true) and is detailed in Appendix A. An example of the segmentation obtained is shown in Figure 6.1, in which each color corresponds to the temporal range of each of the 24 gas sources. This approach increases the spatial range to 8 $\pm$ 3.2 cm (*i.e.* the temporal range to 4 $\pm$ 1.6 sec), so we consider much more points than the simple location of the gas source (which would correspond to some points located around the maximum).

In each temporal range, we then perform a vote between all the labels extracted and associate with the gas source the majority label. This label is then compared to the experimental plan in Figure 6.1.

Evaluation for data from Sniffer arm (Section 6.2.3). In this setup, mixtures occur in different scenarios and it is hard to predict what is occurring outside the gas sources (*i.e.* scent strips). So the spatial range is not increased, and is defined as the spatial area occupied by the scent strips. However, instead of taking a majority label for the whole area as for Sniffer robot, we keep each label $\ell(t)$ for the score. In fact, with Sniffer robot, each gas source is much more away from the other gas sources than with Sniffer arm. In Section 6.2.3, gas sources are so close that it is hard to expect a consensus regarding the labels (this is especially true at the boundaries between 2 gas sources). Let us illustrate the evalutation with an example.

For **Scenario** ②, the Guaïacol source starts from 10.5cm to 15.5cm which corresponds to 25 measurements (speed = 1cm/s and frame rate = 5Hz). For each measurement $\boldsymbol{y}(t)$, we predict the label $\ell(t)$ based on the criterion (6.6). If the label $\ell(t)$ is Guaïacol then the classification is correct, otherwise the prediction is false and so forth for the other gas sources. For **Scenario** ③ (binary mixtures), we consider a success the identification of the full mixture (i.e. not only one VOC among the two).

6.5 Classification of isolated gas sources

In this section, we tackle the issue of classification with the data from Sniffer robot (see Section 6.2.2). As a reminder, these data correspond to 28 laps of a path along which 24 gas sources of 12 VOCs (repeated twice) have been disseminated. The goal is then to apply Algorithm 10 to recognize each one of these gas sources.

6.5.1 Signature extraction and linear model justification

The method used in this chapter is a supervised one, meaning that we need the signatures of the 12 VOCs. To this end, we describe a simple method for estimating them.

According to the model we assume here (eq. (6.1)), the response $\boldsymbol{y}_r(t) \in \mathbb{R}^P$ to a VOC r is simply proportional to its "concentration" $c_r(t) \in \mathbb{R}$. Let us assume that we measure this VOC r during a long period of time, say N_r, at a variable concentration and that we stack these measurements into $\boldsymbol{Y}_r \in \mathbb{R}^{P \times N_r}$. Then, model (6.1) implies that the rank of $\boldsymbol{Y}_r$ is 1. In other words, we can simply write $\boldsymbol{Y}_r = \boldsymbol{k}_r \boldsymbol{c}_r^T$ (in the absence of noise) with $\boldsymbol{k}_r \in \mathbb{R}^P$ and $\boldsymbol{c}_r \in \mathbb{R}^{N_r}$. $\boldsymbol{c}_r$ contains the variations over time and $\boldsymbol{k}_r$ contains the variations across the chemical sensors, in other words $\boldsymbol{k}_r$ is the signature of the VOC r.

Here, we do not measure the VOCs with individual experiments, so we do not have directly a matrix $\boldsymbol{Y}_r$ containing the pure measurements. We extract this matrix for each VOC r by segmenting the signals with a method detailed in Appendix A. With this segmentation, we stack all the peaks corresponding to the VOC r to generate P time series containing "only" the VOC r (assuming that there is no mixture between two cups). Let us assume that these time series are of length N_r with $N_r \geqslant P$, so we have the matrix $\boldsymbol{Y}_r \in \mathbb{R}^{P \times N_r}$. Then, we perform a Singular Value Decomposition (SVD) of this matrix, say $\boldsymbol{Y}_r = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^T$ with $\boldsymbol{U}_r \in \mathbb{R}^{P \times P}$, $\boldsymbol{\Sigma}_r \in \mathbb{R}^{P \times P}$ a diagonal matrix and $\boldsymbol{V}_r \in \mathbb{R}^{N_r \times P}$. Finally, we identify the first column of $\boldsymbol{U}_r$ as the signature $\boldsymbol{k}_r$ of the VOC r and the first column of $\boldsymbol{V}_r$ as the vector $\boldsymbol{c}_r$. Another way could be to perform a Non-negative Matrix Factorization (NMF) instead of the SVD in order to ensure that the coefficients of $\boldsymbol{k}_r$ are non-negative. However, in practice, the coefficients of the extracted $\boldsymbol{k}_r$ are all of the same sign (sometimes negative, so we just flip the sign). All these signatures $\boldsymbol{k}_r$ are then stacked into a matrix $\boldsymbol{K} \in \mathbb{R}^{P \times R}$, which we call the **dictionary**. Each column of this dictionary is a unit-norm vector (*i.e.* $\|\boldsymbol{k}_r\|_2 = 1$).

In order to check whether the rank-1 approximation is correct or not, we compare the matrix $\boldsymbol{Y}_r$ to its best rank-1 approximation $\hat{\boldsymbol{Y}}_r = \sigma_r \boldsymbol{k}_r \boldsymbol{c}_r^T \in \mathbb{R}^{P \times N_r}$. For easier visualization, we just

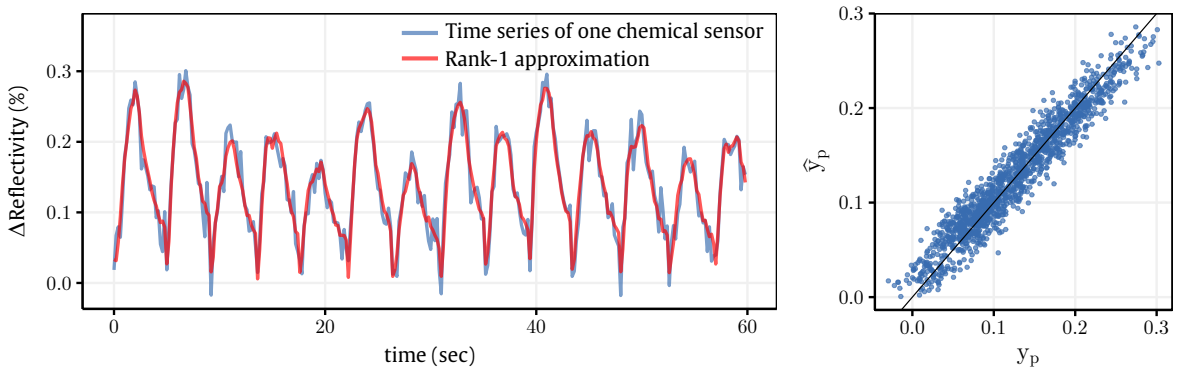


Fig. 6.6.: Justification of the rank-1 approximation. On the left, the blue curve corresponds to the original pure time series of Citral for one chemical sensor and red curve corresponds to its rank-1 approximation. The time series have been truncated to 1 min (out of 4 min), for easier visualization. On the right, another representation (data vs prediction): the more the point cloud is aligned with $y = x$, the better.

represent in Figure 6.6 (left) the time series $\boldsymbol{y}_{p,r}^{\mathrm{T}} \in \mathbb{R}^{N_r}$ (one row of $\boldsymbol{Y}_r$) and its approximation $\hat{\boldsymbol{y}}_{p,r}^{\mathrm{T}} = \sigma_r k_{p,r} \boldsymbol{c}_r^{\mathrm{T}}$ for a given chemical sensor p and a given VOC r (here, Citral).

We also compute the goodness-of-fit R_r^2:

$$R_r^2 = 1 - \frac{\sum_{p,n}(y_{pn,r} - \hat{y}_{pn,r})^2}{\sum_{p,n}(y_{pn,r} - \bar{y}_r)^2} \qquad (6.8)$$

where $y_{pn,r}$ denote the (p,n) entries of $\boldsymbol{Y}_r$ and $\bar{y}_r \in \mathbb{R}$ their average.

R_r^2 equals the proportion of variance explained by the model compared to the total variance in $\boldsymbol{Y}_r$. The values of R_r^2 for each VOC are reported in Table 6.2 and show a generally good fit for all of the VOCs. The closer to 100% R_r^2 is, the better. Both Figure 6.6 and Table 6.2 support the linear approximation, at least in the case of a pure VOC.

VOC	$R^2(\%)$
Butanol	98.0
β-Pinene	99.7
Benzaldehyde	98.0
Citral	92.0
Allyl hexanoate	98.7
Cis-3-hexenol	99.0
Linalool	99.1
Guaïacol	97.9
Geranyl	91.1
Trans-2-octenal	97.9
Acetic acid	94.8
(S)-Limonene	99.7

Tab. 6.2.: The goodness of fit R_r^2 (Eq. (6.8)) for each VOC.

6.5.2 Analysis of the dictionary

The method proposed in this chapter is based on a penalized linear regression. A linear regression can suffer from collinearities or multicollinearites which may exist between the signatures of the matrix $\boldsymbol{K} \in \mathbb{R}^{P \times R}$. A multicollinearity is present when one column of $\boldsymbol{K}$ is equal to a linear combination of the other colums. For non-penalized linear regression, it is easy to show that multicollinearity will cause problems. Indeed, in this case, the classical least squares solution of $\boldsymbol{y} = \boldsymbol{K}\boldsymbol{c}$ is $\hat{\boldsymbol{c}} = (\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{y}$. The inversion of $\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}$ requires that $\mathrm{rank}(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}) = \mathrm{rank}(\boldsymbol{K}) = R$ (assuming $P > R$). This condition is then violated if there exists at least one multicollinearity.

	(S)-Limonene	Citral	Guaïcaol	Cis-3-hexenol
VIF	5.3	10.1	19.7	31.8

Tab. 6.3.: Variance Inflation Factor for 4 VOCs which are used in the setup based on Sniffer arm.

To check for multicollinearity, we use a well-known indicator, namely the Variance Inflation Factor (VIF) [Jam+13]. Considering a signature $\boldsymbol{k}_r \in \mathbb{R}^P$:

$$\text{VIF}_r = \frac{\sum_p (k_{p,r} - \bar{k}_r)^2}{\sum_p (k_{p,r} - \hat{k}_{p,r})^2} \tag{6.9}$$

with $\hat{\boldsymbol{k}}_r = \alpha_0 - \sum_{i \neq r} \alpha_i \boldsymbol{k}_i$ and $\bar{k}_r$ the mean of $\boldsymbol{k}_r$.

Like the R^2 criterion (eq. (6.8)), the VIF is dependent on the notion of variance explained. Here, we regress $\boldsymbol{k}_r$ against the other signatures and the greater the VIF is, the more $\boldsymbol{k}_r$ is linearly dependent on the other signatures. A VIF equals 1 if and only if $\boldsymbol{k}_r$ is linearly independent from the other signatures. A classical rule is that a VIF which is greater than 5 or 10 indicates collinearity problems [Jam+13].

To illustrate the collinearity problem in our case, we focus on a smaller dictionary which will be used for Sniffer arm. The results are reported in Table 6.3. From the Table 6.3, the factors indicate strong collinearities between the signatures even with only 4 VOCs in the dictionary (for the complete dictionary, the results are even worse). These results motivate the use of Algorithm 10 and its ℓ_1-penalty which can help to combat these multicollinearities.

6.5.3 Cross-validation

In order to avoid an overestimation of the score (with the criterion defined in Section 6.4), we perform a cross-validation. In fact, if we extract the signatures with all 28 laps and then predict the labels for these same laps, we may introduce a bias as the training set (extraction of the signatures) is then the same as the testing set.

To perform cross-validation, we divide the 28 laps into β folds (here, $\beta = 5$). The laps corresponding to the first $\beta - 1$ folds are taken for extracting the dictionary and the β^{th} fold is used for testing. Concretely, for the testing laps, we apply Algorithm (10) which returns an intensity vector $\boldsymbol{c}(t)$ for a measurement vector $\boldsymbol{y}(t)$. From this intensity vector, we extract a label for each region of interest which has been previously identified with a segmentation step (see Appendix A). We then repeat the procedure for all the β folds and compute the classification rate and the confusion matrix. Finally, we repeat the entire cross-validation 10 times with new folds. From these 10 cross-validations, the classification rates and the confusion matrices are averaged.

6.5.4 Results

The confusion matrix is reported in Table 6.4 and the average classification score is 73.7% which is much larger than the chance level (8.33%). Despite the difficulty of the task, some

	Truth											
Prediction	Acetic acid	Guaïacol	Benzaldehyde	Linalool	Trans-2-octenal	Allyl hexanoate	Cis-3-hexenol	Butanol	(S)-Limonene	β-Pinene	Geranyl acetate	Citral
Acetic acid	**100**	0	0	0	2	1	0	5	0	3	0	0
Guaïacol	0	**100**	0	0	0	0	0	0	0	0	0	0
Benzaldehyde	0	0	**99**	1	1	1	0	6	0	1	0	4
Linalool	0	0	0	**92**	2	0	1	0	0	0	0	0
Trans-2-octenal	0	0	0	1	**63**	6	0	3	0	4	0	17
Allyl hexanoate	0	0	0	0	18	**72**	0	0	0	1	3	0
Cis-3-hexenol	0	0	0	0	0	1	**59**	16	0	1	0	0
Butanol	0	0	1	1	4	3	40	**59**	2	4	0	16
(S)-Limonene	0	0	0	0	1	5	0	7	**72**	24	9	0
β-Pinene	0	0	0	0	1	6	0	0	20	**48**	3	0
Geranyl acetate	0	0	0	5	1	3	0	1	4	4	**84**	35
Citral	0	0	0	0	7	2	0	1	1	1	0	**28**
None	0	0	0	0	0	0	0	2	0	9	0	0

Tab. 6.4.: Confusion matrix for the data from Sniffer robot. The colored cells correspond to pairs of VOCs which are hard to differentiate. A *None* class has been added when Algorithm (10) did not find any VOC. In fact, it happens that for some regions of interest, Algorithm (10) does not find any VOC (all the intensities are null) whereas the segmentation method did not discard them (especially areas with low signal-to-noise ratio).

VOCs are even perfectly or almost perfectly identified. It means that almost all the gas sources containing these VOCs have been correctly recognized (*e.g.* Acetic acid or Benzaldehyde). However, it happens that Algorithm 10 does not find any VOC (*i.e.* all the VOCs have a null intensity) even if the segmentation does. These estimations have been classified as "*None*" in the confusion matrix and correspond to regions of interest with low signal-to-noise ratio. For these regions, the null solution (*i.e.* $\hat{\boldsymbol{c}}(t) = \mathbf{0}_R$) must have the lower cost compared to solutions with one or more VOCs.

Figure 6.7 represents the distribution of the maximum intensities for each VOC over the time. The diversity of intensities is clear, showing the exhaustion of the liquid phase of some gas sources (*e.g.* β-Pinene) and the stability of others (*e.g.* Guaïacol). It demonstrates that the good classification score cannot be attributed to a single factor such as a simple difference of intensity between VOCs. In fact, if each VOC was responding in a different intensity range, then, the classifier could learn some differences only based on intensities, possibly even after normalization. Here, classification performance can only be explained by the variability in affinities between the sensing materials and the VOCs.

Misclassifications are sometimes due to the fact that some pairs of VOCs are quite hard to differentiate (colored cells in Table 6.4). It is interesting to note that these pairs are sometimes from the same chemical family. For instance (Butanol, Cis-3-Hexenol) are both Alcohols and (β-Pinene, (S)-Limonene) are both Alcenes and they even share the same molar mass. Of course, chemical similarity is not the whole story, since e.g. Linalool is not confused with other Alcohols. The misclassifications between Citral and Geranyl acetate can be attributed to the lower signal-to-noise ratio for these two VOCs. This low signal-to-noise ratio can be explained by both their low volatility and their low affinity with the sensing materials.

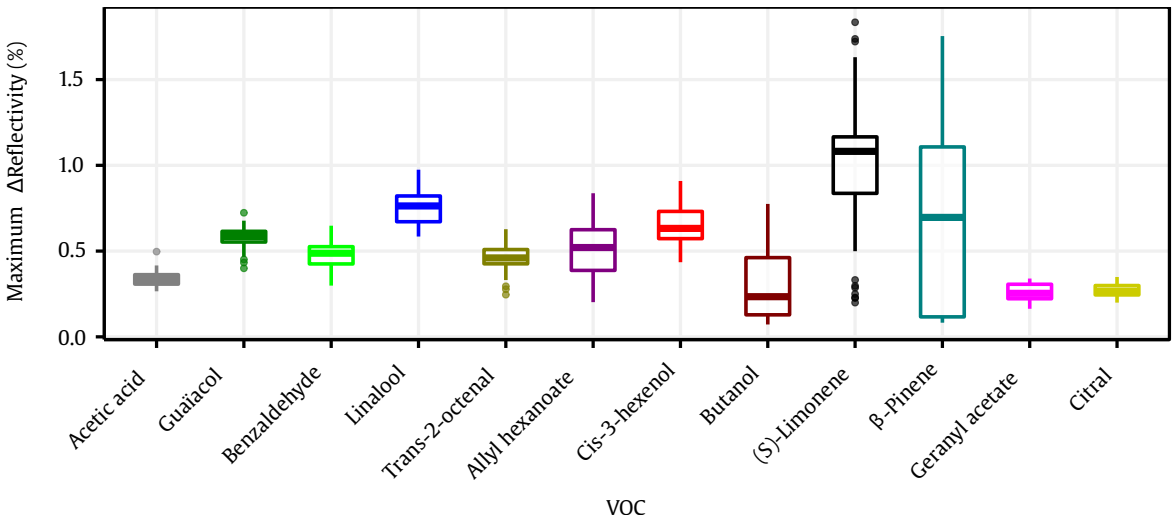


Fig. 6.7.: Distribution of the maximum intensities over the laps for each VOC.

6.6 Gas unmixing

In this section, we tackle the issue of gas unmixing with the data from Sniffer arm (see Section 6.2.3). We remind the reader that we have generated various realistic scenarios with increasing complexity: ① the gas sources are isolated and spatially close, ② the gas sources are successive trails of pure compounds and ③ the gas sources are successive trails of binary compounds. In each scenario, the data correspond to 20 sweeps (from left to right, the results for the other direction are reported in Appendix C.2). Each one of these 20 sweeps is then processed separately for the unmixing with the previously generated dictionary and with Algorithm 10.

6.6.1 Building and pruning of the dictionary

The dictionary $\boldsymbol{K} \in \mathbb{R}^{P \times R}$ is built with the method detailed in Section 6.5.1 and from the whole data set from Sniffer robot. This last point is an important aspect of the following results: it implies that the training setup is not the same as the testing setup. This characteristic is quite important in practice. Indeed, the dictionary will be always generated in a separated setup in order to be used in the field afterwards. So, training with Sniffer robot and testing with Sniffer arm will assess the robustness of the proposed method.

From the dictionary, we first extract the set of the 4 VOCs which are present in the scenarios ① to ③, namely: Citral, (S)-Limonene, Guaïacol and Cis-3-hexenol. At least one VOC is always used as a control, meaning that at least one VOC is present in the dictionary but **not** in the experiment. We call the VOCs actually present the **target VOCs**. We expect that the estimated intensity of the **control VOC** will be close to or equal to 0. This is a key point of our results; indeed, an eNose is a non-specific device that can generate signatures for a broad variety of VOCs, contrary to specific sensors which are designed for one or two VOCs. In practice, an eNose may be less effective than specialized sensors if we only target 2 specific VOCs. So there is no reason to favor an eNose except if we want to use it for a large amount of

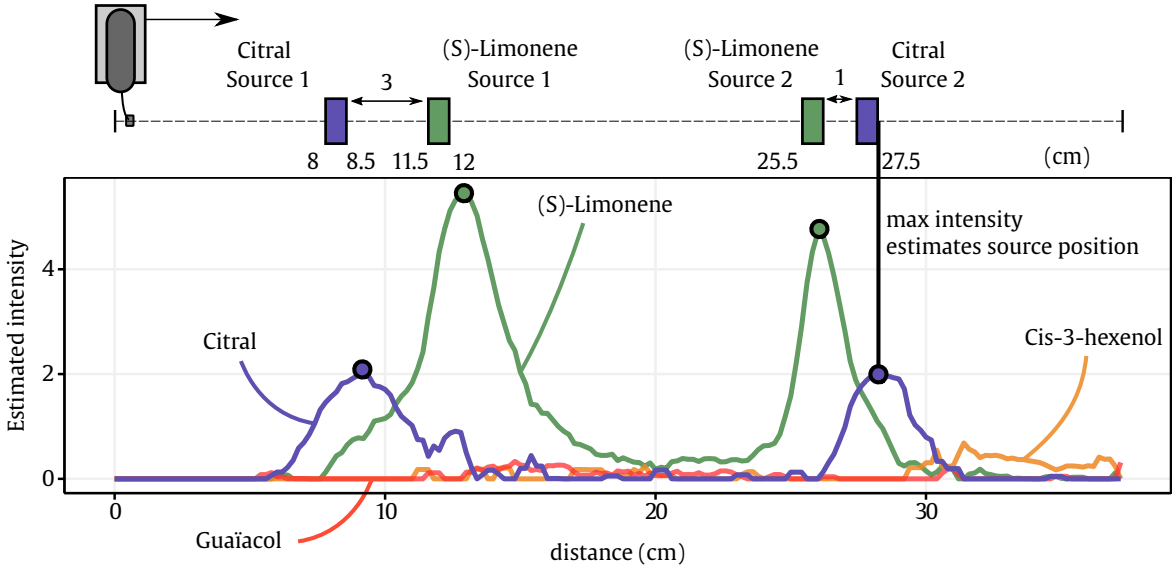
applications with many different VOCs. That is why, in practice, the dictionary will be as large as possible and only an unknown subset of VOCs from the dictionary will be relevant for a given application. To our knowledge, by considering a larger dictionary than the number of VOCs actually present, we go further than most of the studies in the field and especially studies which perform mixtures-learning. Furthermore, these 4 VOCs are clearly not the best subset of the 12 VOCs acquired with Sniffer robot. Indeed, by looking at the confusion matrix in Table 6.4, the subset containing Acetic acid, Guaïacol, Benzaldehyde and Linalool would probably lead to much better results than the ones presented here.

The results with the dictionary of 4 VOCs are described in the next section. Afterwards, we progressively extend the dictionary by adding one VOC at a time. The order of the VOCs is defined with the help of the confusion matrix (Table 6.4). Given the current set of VOCs in the sub-dictionary, we simply add all the confusions made with these VOCs (*i.e.* we add amongst themselves the columns of the confusion matrix containing the set). Then, we take as new member of the set the VOC that has the lowest confusion coefficient. With this method, the order of the VOCs (after the 4 VOCs already chosen by default) is: Acetic Acid, Allyl-hexanoate, Linalool, Benzaldehyde, Trans-2-octenal, β-Pinene, Geranyl acetate and Butanol. The task is then harder and harder and, at the end, the whole dictionary is used in the unmixing.

6.6.2 Results with a dictionary of 4 VOCs

The intensities estimated for the 4 VOCs in the 3 scenarios are represented in Figure 6.8. Each color stands for a VOC and each line corresponds to a sweep (a color gradation indicates the sweep index).

For **Scenario** ①, the intensities are reported in the top panel of Figure 6.8. The intensities of Citral and (S)-Limonene clearly indicate good unmixing from the signals previously shown in Figure 6.2b. We can notice that the intensities of the controls (namely Cis-3-hexenol and Guaïacol) are not strictly null. However, these intensities correspond mainly to transition areas (from one gas source to another) or to areas with low signal to noise ratio (SNR). In addition, the intensities of the controls are much smaller than the intensities of the two VOCs actually present. By computing the classification criterion defined in Section 6.4, we reach a noteworthy score of 90% (98% for the other direction) for the spatial areas defined by the scent strips (corresponding to a total of 240 measurements). In this scenario, we even go further by estimating the location of each isolated gas source. For that, we simply take as the position of a gas source, the position of its maximum intensity. The average estimated locations for the 20 sweeps are reported in Table 6.5. These results highlight a good estimation of the position of the gas sources when they are far enough from each other (separated by 3cm). The same task is harder for a smaller distance (1cm) for which the distance is overestimated. In addition, (S)-Limonene presents better estimation results compared to Citral. This can be explained by the possible lags introduced in Section 6.4, especially the chemical lag and the lag related to the airflow. Indeed, Citral is a heavier



	Citral		(S)-Limonene		Distance	
	Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
Mean ± Stand. dev. (cm)	8.6 ± 0.35	27.9 ± 0.50	11.9 ± 0.31	25.5 ± 0.28	3.3 ± 0.41	2.3 ± 0.47
Ground truth (cm)	[8,8.5]	[27,27.5]	[11.5,12]	[25.5,26]	3	1

Tab. 6.5.: Estimation of the position of the isolated gas sources from Scenario ① with the default dictionary (Citral, (S)-Limonene, Guaïacol, Cis-3-hexenol), based on the location of the maximum intensity. The column distance refers to the spatial distance between Citral and (S)-Limonene. The two numbers of the ground truth correspond to the beginning and to the end of the scent strip (*i.e.* the gas source) along the path. Top figure gives a graphical explanation of the procedure for estimating the gas source position from unmixing results. Top figure corresponds to the unmixing of a single sweep (in Figure 6.8 all the sweeps are represented), based on a dictionary of 4 VOCs.

molecule than (S)-Limonene (see Table 6.1) so its transport and its interaction with the chemical sensors can take more time and thus delay the measurement.

For **Scenario** ②, the results are represented in the middle panel of Figure 6.8. Again, the estimated intensities correspond to the spatial areas of the gas sources. However, the estimated intensity of the control (here, Cis-3-hexenol) is no more negligible and no more restricted to low-SNR regions. We notice that the intensity of the control VOC depends mainly on the VOC present. Indeed, the gas source containing Guaïacol is well estimated whereas the estimations for the other gas sources ((S)-Limonene and Citral) are more affected by the presence of Cis-3-hexenol in the dictionary. This observation is a consequence of the existence of linear dependencies between the signatures. Despite these correlations, we find a classification score of 79% (92% for the other direction), out of 1,520 measurements. This score is quite good in view of the difficulty of the task. Even if the intensity of the control VOC is high, it is still lower than those of the target VOCs.

Finally, for **Scenario** ③, the intensities are reported on the bottom panel of Figure 6.8. At first sight, the results seem better than for **Scenario** ② which is simpler, especially if we focus on the control VOC. In fact, the control VOC is no more Cis-3-hexenol but now Guaïacol.

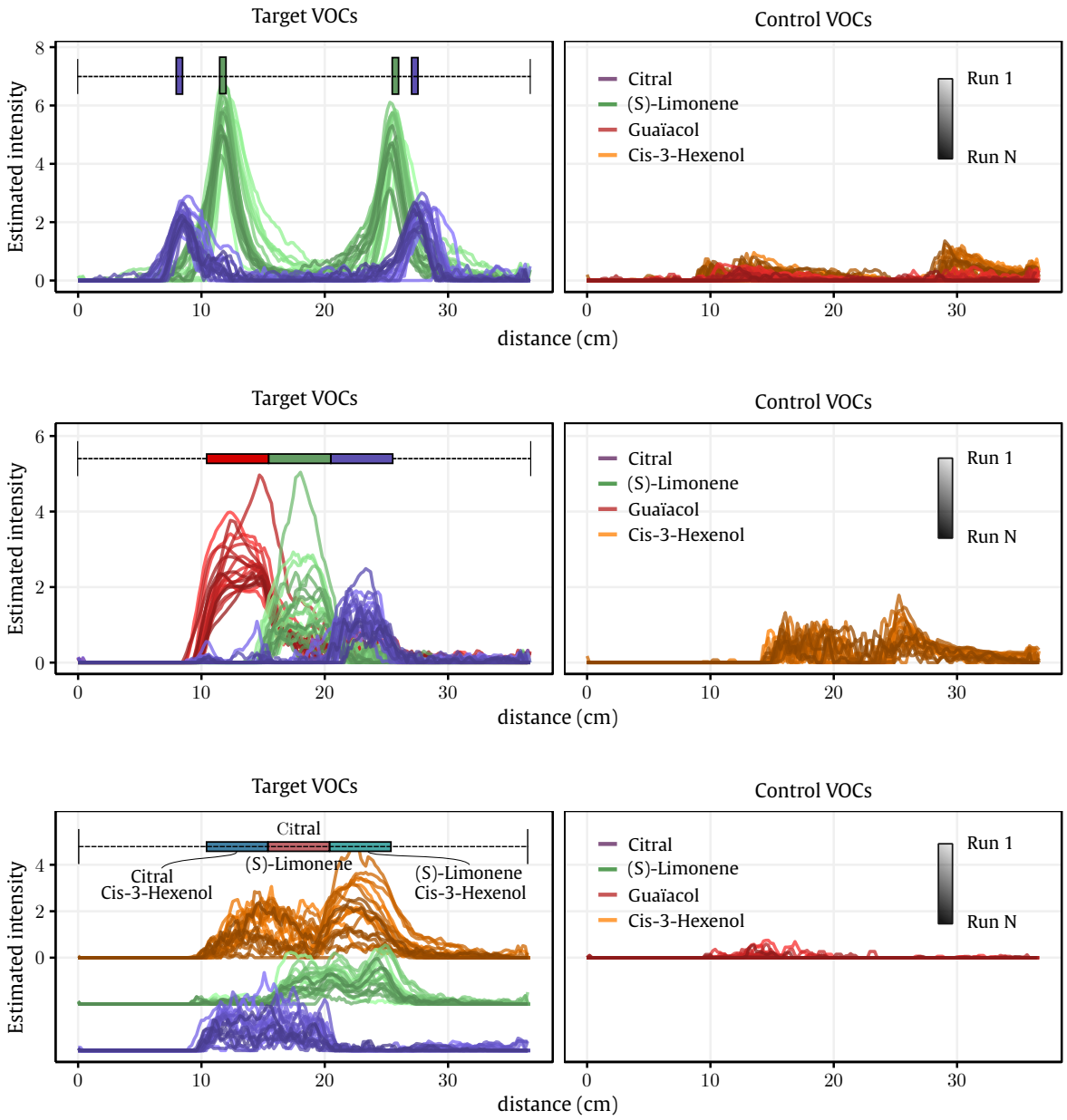


Fig. 6.8.: Results of the proposed algorithm for the different scenarios introduced in Section 6.2.3. Top figures: Scenario ①. Middle figures: Scenario ②. Bottom figures: Scenario ③. Each line corresponds to one sweep (first sweep is the lighter). Each color corresponds to the estimated intensity of the given VOC at the distance d. The results have been generated with the default dictionary ((S)-Limonene, Citral, Guaïacol, Cis-3-hexenol). For Scenario ③, the intensities of Citral and (S)-Limonene have been vertically shifted for easier visualization.

Guaïacol seems to be a much better control VOC than Cis-3-hexenol, since its presence in the dictionary has little impact on the unmixing. Indeed, the estimated intensities of Guaïacol are close to 0, as expected. This can be explained by a VIF which is lower for Guaïacol (see Table 6.3) and by the Table 6.4 which shows that Guaïcaol is perfectly identified despite the 11 other VOCs. For the present VOCs, the estimated intensities match the location of the gas sources, especially for the gas sources containing Citral and (S)-Limonene. The intensities of Cis-3-hexenol are less simple to analyze. Indeed, the estimated intensities indicate the presence of Cis-3-hexenol in the middle gas source while Cis-3-hexenol is not present in

this gas source. We explain that by the presence of ternary mixtures in combination with the desorption of Cis-3-hexenol. A comforting fact is that these intensities clearly decrease when the eNose goes over a gas source which does not contain Cis-3-hexenol. The score reaches 70% (69% in the other sweep direction) which means that the identity of 70% of all binary mixtures (corresponding to 1,520 measurements, i.e. 1,520 mixtures) has been well predicted.

6.6.3 Results with a dictionary of increasing size

The previous results emphasize that we can achieve good classification performance even if one or two VOCs are absent in the experiment but present in the dictionary. Here, we go even further by adding one-by-one each VOC from the full dictionary. As a reminder (see Section 6.6.1 for details), the order according to which the VOCs are added, is the following: Acetic acid, Allyl-hexanoate, Linalool, Benzaldehyde, Trans-2-octenal, β-Pinene, Geranyl acetate and Butanol.

Concretely, we start by adding the signature of Acetic acid to the previous dictionary of size $R = 4$ ((S)-Limonene, Citral, Guaïacol, Cis-3-hexenol). Then, we apply Algorithm (10) to unmix the signals and estimate the intensity of each VOC of this new sub-dictionary. Ideally, the estimated intensity of Acetic acid is close to 0. However, in practice the new VOC could considerably disrupt the unmixing, especially due to correlations which may exist between this new signature and the signatures already present. To assess the influence of this new VOC, we simply generate the label $\ell(t)$ based on the maximum intensity of $\boldsymbol{c}(t)$ (see Section 6.4 for details). This predicted label is then compared to the position of the gas sources. Afterwards, we repeat the procedure by adding Allyl-hexanoate and so forth for the others. Until we reach the upper limit of $R = 12$ VOCs (the full dictionary), the task becomes harder and harder with each new VOC we add. In Figure 6.9, we report the average score for each dictionary size and each scenario. In Figure 6.10, we represent the spatial distribution of the classification score for each dictionary size and each scenario. Figure 6.10 highlights which locations along the path cause most misclassifications.

From Figure 6.10, for **Scenarios** ① and ②, we see that the scores start to considerably decrease only at $R = 10$ VOCs. It is noteworthy that even with 9 VOCs in the dictionary, we can reach a classification score of 88% for **Scenario** ① and 74% for **Scenario** ②. In fact, $R = 10$ corresponds to the addition of β-Pinene which is highly-correlated (99.8 %) to (S)-Limonene (for the two scenarios, the gas sources of (S)-Limonene are no longer identified). Afterwards, at $R = 11$, the addition of Geranyl acetate produces another large decrease of the scores (again, for the two scenarios, the gas sources of Citral are no longer identified). In fact, the confusion matrix (see Table 6.4) already showed that Geranyl acetate can cause a lot of misclassifications for Citral. At $R = 11$ or $R = 12$ (the full dictionary), the final classification scores reach 2% for **Scenario** ① and 49% for **Scenario** ②. This shows that the addition of Butanol has no effect on the classification. The difference between **Scenario** ① and ② is explained by the gas source containing Guaïacol which is present in **Scenario** ② but not in **Scenario** ①. The correct identification of Guaïacol agrees with Table 6.4, in which Guaïacol

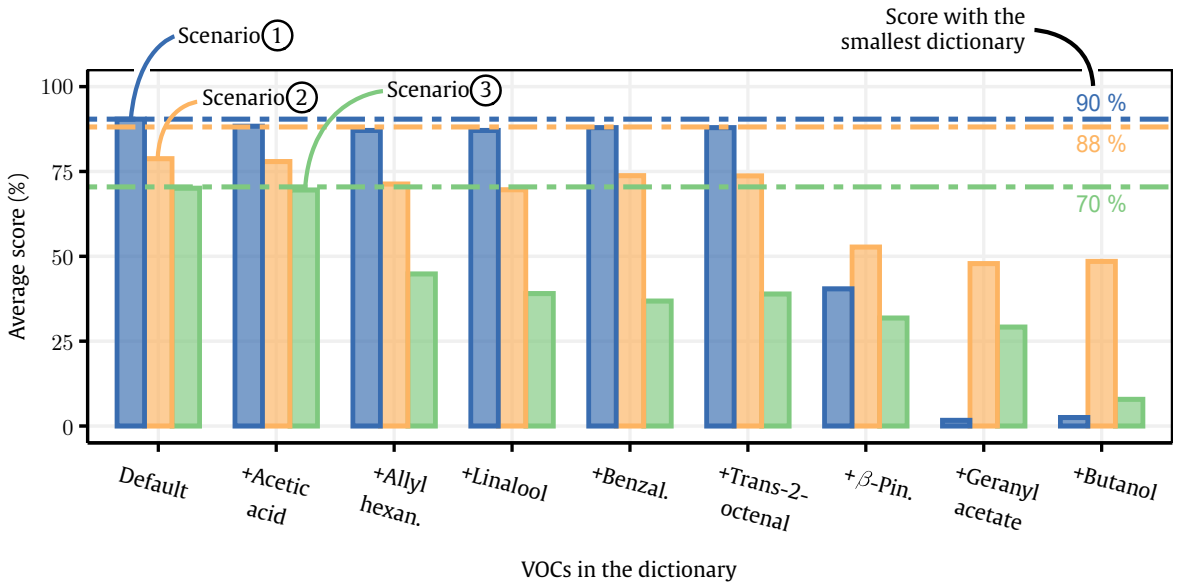


Fig. 6.9.: Average score in each scenario and for each dictionary (by averaging over space and across the 20 sweeps). A dashed line indicates the score obtained with the smallest dictionary (*e.g.* for **Scenario** ① the smallest dictionary is (Citral, (S)-Limonene)). The default dictionary corresponds to (Citral, (S)-Limonene, Guaïacol, Cis-3-hexenol).

is also reliably identified. Finally, the hardest areas to identify are the transition areas. This observation is really clear in **Scenario** ② and reflects a drawback of our assessment method. Indeed, the score is based on the spatial position of the gas sources but it does not take into account the diffusion of the VOC or some interactions which can occur between different VOCs. For example, there is no reason that the gas source of (S)-Limonene for **Scenario** ② starts exactly at the beginning of its scent strip and not a little time after or before due to diffusion or due to interaction with Guaïacol. So our assessment method may be biased to some extent.

In **Scenario** ③, things are even more complicated. If we only refer to the scores, they "continuously" decrease starting with the introduction of Allyl-hexanoate at $R = 6$. It means that when we add two control VOCs (Guaïacol and Acetic acid), there is no strong effect on classification performance compared to the best dictionary ((S)-Limonene, Citral, Cis-3-hexenol). In this case, most of the misclassifications are due to the middle gas source in which the intensity of Cis-3-hexenol seems to be over-estimated (see an example in Figure 6.8). However, again, the score computed with the spatial position of the gas sources does not take into account all the possible variations which can occur. Here, it is likely that ternary mixtures occur around the middle gas source (Cis-3-hexenol is present in the two extreme binary trails). Ternary mixtures are not taken into account by the criterion and a single misclassification leads to a global misclassification of the mixture according to our criterion (recall that a correct classification means that the 2 VOCs present in the mixture are well identified, and not just one over the two). This explains why the scores for Scenario ③ are worse and decrease faster than in Scenarios ① and ②.

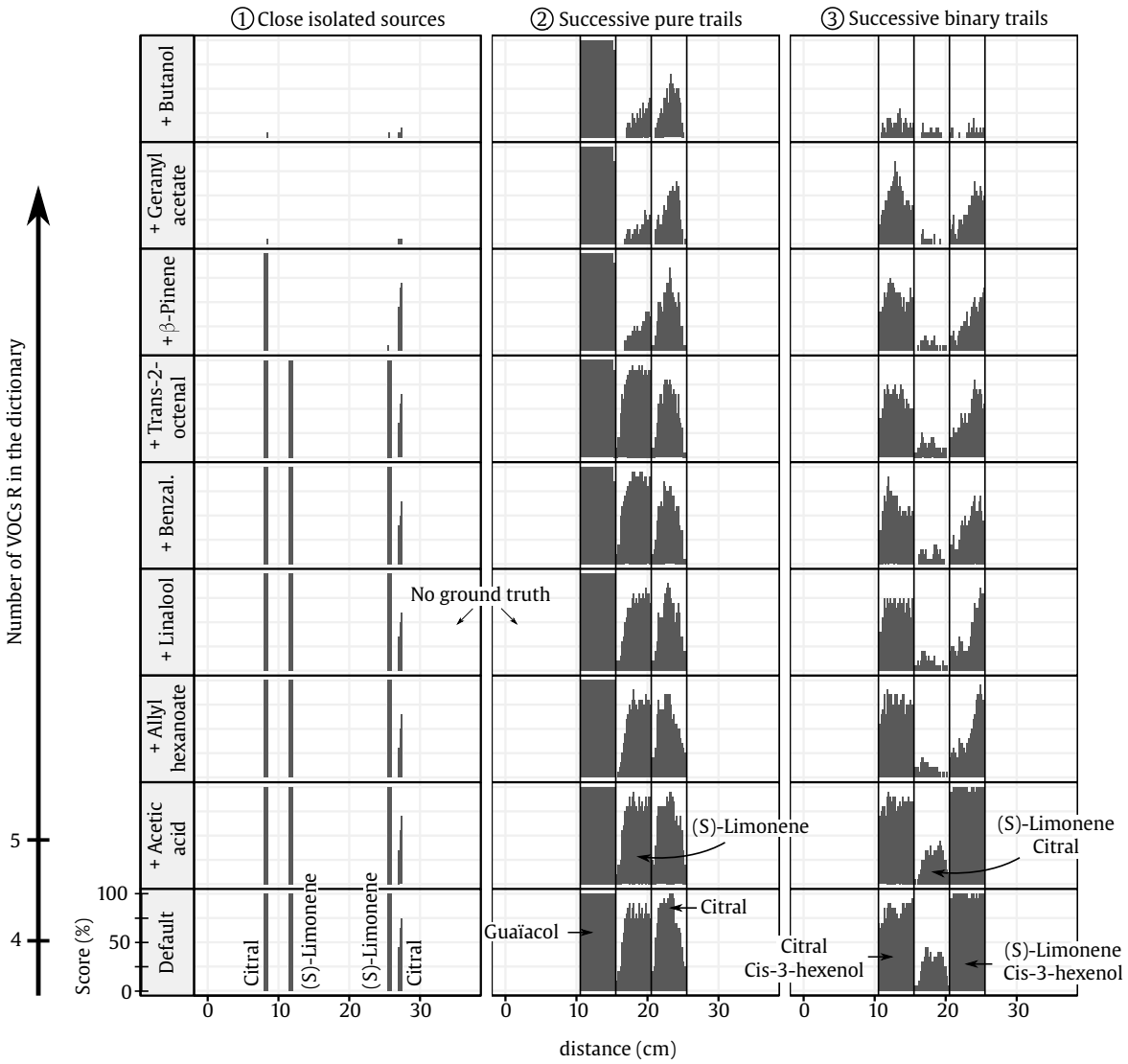


Fig. 6.10.: Influence of the size of the dictionary on spatial distribution of the classification score. From the bottom up, the number of VOCs in the dictionary is increasing, starting from the default dictionary ((S)-Limonene, Citral, Guaïacol, Cis-3-hexenol). Left to right, each panel corresponds to the spatial distribution of the score for each scenario (for a position d, the score is the average across the 20 sweeps). The spatial distribution highlights the misclassifications (mainly in transition areas) in each scenario and for each dictionary. For **Scenarios** ② and ③, a black line indicates the theoretical transitions between one gas source to another.

6.6.4 Discussion

The unmixing results seem quite acceptable in all scenarios. To our knowledge, it is the first demonstration of such unmixing results with an electronic nose in an uncontrolled environment with open sampling. Moreover, these results have four additional key aspects: first, the mixtures have not been previously trained (only the dictionary of pure compounds and, on top of that, with a separated setup), second, the number of VOCs in the mixtures is at no time assumed known during the unmixing process, third, the dictionary is much larger than the number of VOCs actually present, and finally, all the results have been generated in a real-time fashion.

However, the results are obviously perfectible. First, they can be affected by the correlations or the linear dependencies which exist between the signatures of the dictionary. The ℓ_1-penalty of Algorithm 10 improves the solution in the presence of collinearity compared to classical least squares solutions. But it improves the solution in a way which can be unacceptable in our case. Indeed, if two signatures are strongly correlated then the Lasso will tend to randomly select one of the two signatures, putting the other to zero [Tib11]. In other words, in the field, if only the VOC A is present and strongly correlated to a VOC B present in the dictionary, then the Lasso solution may indicate the presence of sometimes A, sometimes B. If A is a harmful VOC and B is a safe one, it is easy to understand the consequences of such an error. Improvements can be made by decorrelating A from B or by simply discarding the signature of B from the dictionary. Another line of research to differentiate A and B, is also the design of additional sensing materials. Indeed, if A and B have currently correlated signatures, it means that the actual sensing materials are not enough discriminative regarding these two compounds. In other words, it means that the sensing materials interact currently in a similar way with A and B.

Second, even if the linear model seems a good approximation here, a non-linear relationship exists certainly and may improve the unmixing. A non-linear model motivated by theory would be hardly achievable due to the variations in multiple crucial parameters during an experiment, such as humidity, temperature or concentration of the VOC. An empirical improvement could be to add an interaction term such the one proposed by [Llo+98] for the MOS sensors. In our notation, the model would correspond to $\boldsymbol{y} = \sum_r (\boldsymbol{k}_r c_r - \sum_{i \neq r} \boldsymbol{a}_{ir} c_i c_r)$ where $\boldsymbol{a}_{ir} \in \mathbb{R}^P$ is the interaction vector between the VOC i and the VOC r. This interaction term could then capture some variations which are not taken into account by the linear model. However, some tradeoffs are involved since a more complicated model can require more experiments (for calibrating some parameters for instance) and can also be much more complicated to fit due to the existence of local minima, indeterminations, or numerical unstability.

6.7 Conclusion and future works

This chapter reports two main results regarding the use of Aryballe's eNose for a robot application. First, we succeeded in recognizing up to ~73% of 24 isolated gas sources containing 12 different VOCs. Second, we succeeded in unmixing binary and ternary mixtures occurring in various realistic scenarios. When gas sources are close enough for significant mixing to take place, we reach a classification score of 90%. When the robot goes over several successive trails of pure compounds, we reach a score of 88%. Finally, when the robot goes over several successive trails of binary compounds, we reach a score of 70%.

Our methodology has several noteworthy features. First, the results have been obtained with a single, unified, processing pipeline. Second, this pipeline is based on the use of a dictionary of pure VOCs, which is quick to generate. Third, real-time estimation of the intensities does not require any information about the number of VOCs present in the mixtures. Finally, the

dictionary was generated in a different setup and was always larger (and sometimes much larger) than the number of VOCs currently present.

Future works. The experiments we have carried out can be extended in many ways. First, besides the convection induced by the funnel, there is not so much advection in our experiments because we are in a closed room. To be even more realistic, further studies can incorporate some wind, but wind can decrease recognition performance (see the study of [Ver+13]). Second, the actual concentration of each VOC is unknown in our experiments. This was a restrictive factor for the assessment of our results and it should be removed in further studies. The concentration would be especially highly valuable in order to assess if the intensity matrix that we estimate is related in some way to the true concentration of the VOCs. For instance, the use of a PhotoIonization Detector can provide such information (see [HB+14]) but at the cost of a high airflow (300-500 mL/min) which could easily impoverish the gas mixture. Third, the scale of the setups that we have built can be considered small compared to other setups in the literature (see [Fan+19]). Even if the small scale used here was appropriate for assessing our algorithm, the next setup must be larger to gain even more realism. Finally, the experiments were carried out on the same day or with a one-day interval. In practice, the dictionary is generated once and has to be used several days, weeks or even months after its generation. It could then be interesting and highly valuable to reproduce the results of Chapter 4 in the more complicated setting used here.

Regarding the algorithms that we have proposed, the development of methods for decorrelating the signatures and the enrichment of the mixing model have been already discussed in Section 6.6.4. However, other improvements can be imagined. First, if we try to extend the results of Chapter 4 on drift correction, we will need to design a method to compensate for the drift (see the review of [MGG12]). In that case, the signatures of the dictionary will vary and these variations will lead to poor results if no correction is applied. Another solution in the presence of drift would also be to regenerate the dictionary, but it would be inconvenient in practice. Second, the measurement is affected by temporal lags, related for instance to the chemical reactions between the VOC and the sensing materials (see Section 6.4 for details). In source localization, these lags can impair the estimation of the position of the gas source. Deconvolution techniques can then be deployed to mitigate the effects of the lags (see [MBM19]). For gas mixtures, these techniques can be particularly difficult to develop since the chemical lag depends strongly on the VOCs in the mixture, which are unknown. Finally, an interesting extension of our results would be to consider the blind case, in other words the Blind Source Separation (BSS) framework. The main difference with our study is that we would no longer assume knowledge of the signatures matrix $\boldsymbol{K}$ for the unmixing. Only very few recent studies have dealt with this issue (*e.g.* [MBC18], [Mad+18]) but a BSS method could considerably reduce the experimental time required for learning.

An opening towards GC-eNose. In Chapter 1, we introduced Gas Chromatography as the gold standard for separating a gas mixture into its individual components (see Section 1.1.2). The working principle of GC relies on the use of a column whose internal surface is covered

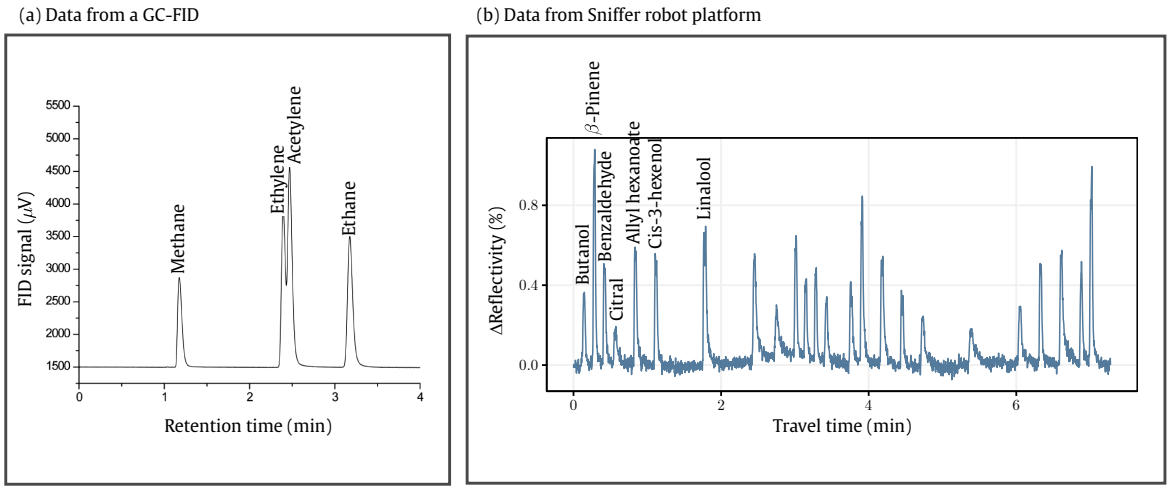


Fig. 6.11.: Comparison between the data from a GC-FID (a) and the data from our robotic setup (b). x-axis corresponds to time but time has not the same meaning for the two figures: in (a), it corresponds to the time spent by each VOC in the GC column, and in (b), it is related to the distance followed by the robot (which then pass over gas sources). However, it is interesting to draw a parallel between these two figures since data look alike. Figure (a) was retaken from [LT11].

with a molecule. When the gas mixture is injected in this column using a carrier gas, each component of the mixture interacts differently with the molecule covering the internal surface of the column and with the carrier gas. Depending on these interactions, each VOC composing the gas mixture will leave the column at a different time, called retention time. This process allows the gas mixture to be decomposed into its individual components over time. However, GC does not allow to detect, identify or quantify the VOCs which leave the column. GC is therefore always used in combination with another instrument serving as detector (see Section 1.1 for the review of some detectors).

In Figure 6.11, we compare the data obtained from a GC-FID (figure retaken from [LT11]) and the data acquired after one lap with Sniffer robot. We can notice how similar data are. This comparison paves the way to an interesting opening regarding the results highlighted in this chapter: the combination GC-eNose. Aryballe's eNose would then be used as detector in the output of a GC, instead of using a mass-spectrometer or a flame ionization detector. The response would probably look alike Figure 6.11b, meaning a series of peaks appearing over time. From our study, we know that it would be possible to identify each peak, so it would be possible to identify each VOC which is leaving the column (on top of that in real-time). For instance, a FID used in combination with a GC is not capable of such a feature. From our unmixing results, we also know that it would be possible to deal with unresolved peaks that one sometimes observed with a GC. In Figure 6.11a, the peaks of Ethylene and Acetylene illustrate this issue [LT11]. However, we still do not know if quantification would be feasible.

We can list, at least, 3 further studies which could be performed based on this remark. First, we could compare the results of a GC-FID with the results of the GC-eNose. This comparison would allow to know if the peaks observed with our robotic setup (see Figure 6.11b) could be integrated to estimate the VOC concentration, just like what it is classically done with a

FID signal. This comparison would also allow to know if the estimated intensities are directly related to VOC concentration. Second, with a GC-eNose approach, we could analyze much more complex mixtures composed of several dozens of different VOCs. For the moment, this task would be quite hard for Aryballe's eNose (and for any eNose) since the instrument boards "only" ~20 different sensing materials. In fact, our unmixing results show that even the unmixing of ternary mixtures is not trivial. In this case, GC would help to physically separate the components of the mixture over time. In the same vein, a third utility of the combination GC-eNose is to simply increase the selectivity of the eNose. A good example is the study of Zampolli et al. [Zam+05]. In their study, selectivity is crucial for monitoring the concentration of pollutants in the air. In this case, non-harmful VOCs can be present at high concentrations while toxic compounds are present at much lower concentrations [Zam+05]. The response of the instrument to non-harmful VOCs may therefore hide the response to toxic VOCs. To alleviate this issue, Zampolli et al. used a MOS-based system in combination with a GC. By separating the different VOCs with GC, it allows to increase the selectivity towards harmful VOCs. To reach to these results, they also built a micromachined GC column [Zam+05]. In fact, in Chapter 1, we stressed that GC is not suitable for measurements in the field. To that end, GC needs to be miniaturized to be used in combination with an eNose (for example, by considerably reducing the length of the column).

7 Conclusion

An electronic nose takes inspiration from the main steps of natural olfaction. Their working principle is based on the use of an array of non-specific chemical sensors. This non-specificity allows the instrument to interact with a broad variety of volatile compounds, making it potentially useful for a large range of different applications. It actually concerns any application where a compound is released in the air. To name a few examples, an electronic nose could be useful for monitoring pollutants, analyzing exhaled breath or detecting gas leaks.

Despite all these possible applications, these instruments are not a widespread technology and show good results in the lab but they often never go beyond. Several reasons can be pointed out such as the lack of repeatability of the instrument over time, the low statistical reliability of the studies, the small number of sensors typically used, etc...

By using optical transduction methods, the number of sensors can be considerably increased without increasing the system complexity. In this thesis, we used a new instrument developed by Aryballe (Grenoble, France) and we proved several important results, both theoretical and experimental.

First, we characterized the response of the instrument in the presence of a pure compound. The model is based on the combination of the Langmuir model, for describing the chemical reactions, with a linear approximation of the transduction method. We showed that the general form of the model fits well real data. However, we found that the features extracted from the model were less discriminative than other more classical features, showing the little interest of model-based features for classification purpose. But modeling is not just about feature extraction for Machine Learning algorithms and is also crucial in Source Separation. This field aims at estimating individual responses and/or concentrations from several measurements of a gas mixture. When one works with non-specific chemical sensors, Source Separation methods are particularly relevant for a simple reason: the response of a non-specific chemical sensor to a gas mixture is the result of a combination between individual responses (of each compound) and concentrations. To that end, we derived a non-linear mixture model relating the response of the instrument to the concentration of each compound and each individual response. To our knowledge, this model is new so we provided a theoretical analysis of some of its properties. Then, we developed an inversion algorithm based on Alternating Least Squares to estimate both concentrations and individual responses from a measurement matrix. The proposed algorithm showed good results in simulation. However, the model still needs to be validated on real data.

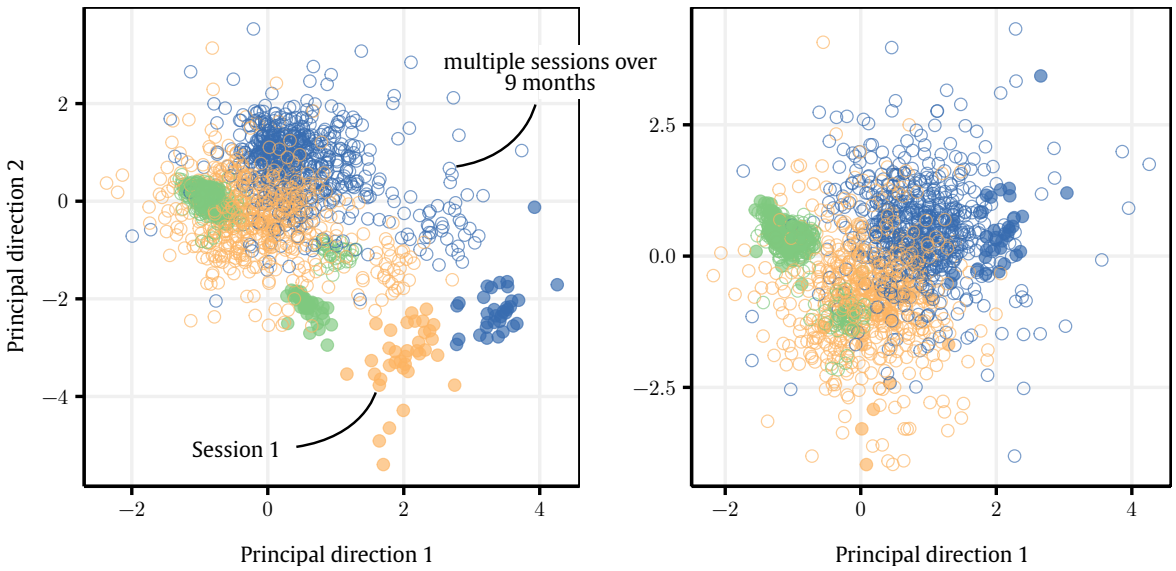


Fig. 7.1.: Projection of the data set onto the two first Principal directions. This data set has been acquired during the thesis for studying the drift issue and corresponds to multiple measurement sessions acquired over 9 months. Each color stands for a volatile compound. Left figure corresponds to the raw measurements: we cannot generalize from Session 1 to other sessions. Right figure corresponds to the measurements corrected with one of our two proposed methods (here, a method that we called EMC^2). The improvement is clear, even if the correction is not perfect (we can see that one session of the "green class" has been poorly corrected).

Second, all types of chemical sensors are prone to drift over time, meaning that their response to a given compound changes. This drawback causes a serious problem for Machine Learning algorithms: what is learnt during a day d may not be generalized to a day $d + 1$ (it can be seen as a special case of "Transfer Learning" in Machine Learning). We studied the drift issue during the thesis by building an extensive data set composed of multiple measurement sessions over 9 months. We showed that Aryballe's instrument is no exception to the rule: if we use the first session as training then the classification score of subsequent sessions rapidly decreases over time. To correct this drift, we proposed two new algorithms. These two methods are based on the assumption that the drift follows a preferred direction but they assume nothing else than the labels of Session 1. This is a significant improvement compared to most of the methods in the literature which assume the existence of some labelled measurements in the subsequent sessions. In Figure 7.1, we represent the data set projected onto the two first Principal directions of the data from Session 1. The improvement is clear.

Third, we studied and showed the stereoselectivity of the instrument. Stereoselectivity is the ability of a system to tell two mirror molecules apart. These two mirror molecules are called enantiomers (an example of two enantiomers are left and right hands) and have almost identical physico-chemical properties (mass, solubility, volatility, boiling point, etc...). This last point makes their discrimination as one of the hardest challenges in analytical chemistry. A surprising fact is that natural olfaction can actually differentiate between two enantiomers.

An even more surprising fact is that Aryballe's instrument has also this ability. We studied the stereoselectivity of the instrument for two different pairs of enantiomers. Over two long measurement sessions, the instrument was not only capable of a perfect discrimination inside the same session but was also able to learn from a session and perfectly generalize to another (see Figure 7.2). We showed that this result is supported by the peptidic nature of the chemical sensors.

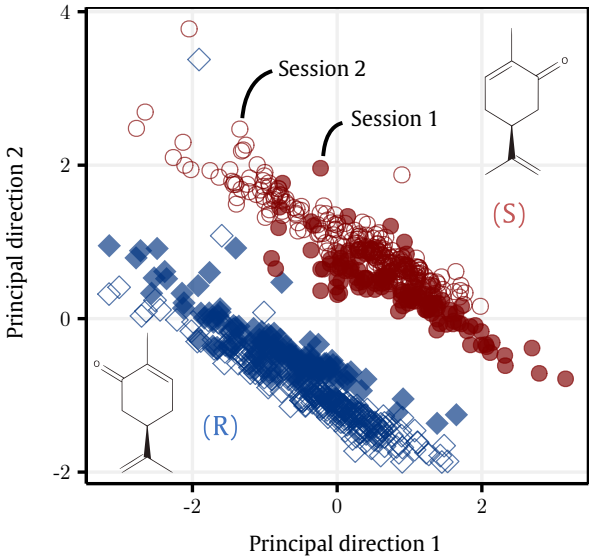


Fig. 7.2.: Projection of the data set onto the two first Principal directions. Over two long sessions, we showed that the instrument is able to discriminate the two enantiomers of Carvone (one is called (R) and the other (S)). Similar results were obtained for the enantiomers of Limonene.

Last but not least, we contributed in many ways to robot olfaction by designing new algorithms and building different experimental setups during the thesis. For monitoring environmental pollution, an electronic nose can be static, waiting for compounds to reach by themselves the instrument. But, other applications, such as gas leak detection, require in addition that the instrument can move and locate in space the gas source. In the literature, this field of applications is referred to as robot olfaction. In this context, we proposed an algorithm which is capable in real-time of both identifying a gas source and separating gas mixtures. The algorithm is based on the existence of a known dictionary containing the individual responses of each compound. It assumes a linear model relating the response of the instrument to the gas concentrations. By using this algorithm, we first succeeded in well identifying 24 gas sources of 12 different compounds which were disseminated in the environment. To fully appreciate this result and to give a general idea, a great number of studies in the robot olfaction field restrict the case to two gas sources. Then, we investigated an issue which is usually ignored by the literature in this complicated (but realistic) context, namely Source Separation. In the field, gas mixtures can easily occur either due to spatially close gas sources or due to the fact that the source is actually a mixture. We studied these two possible cases in three different robotic scenarios. They differ on difficulty but they are all realistic scenarios which one can encounter in the field. We demonstrated that the proposed algorithm can unmix these signals in real-time, thus telling us where each compound is located along the path even in the presence of gas mixtures. One example of these noteworthy results is reported in Figure 7.3. We also investigated the influence of the size of the dictionary on the unmixing performance, by considering up to 12 differents compounds in the dictionary (whereas only three compounds were actually present in the experiments). This issue is even more ignored by the literature and, to our knowledge, it is the first time that such a study is conducted with an electronic nose. Our results show that the larger the dictionary, the harder

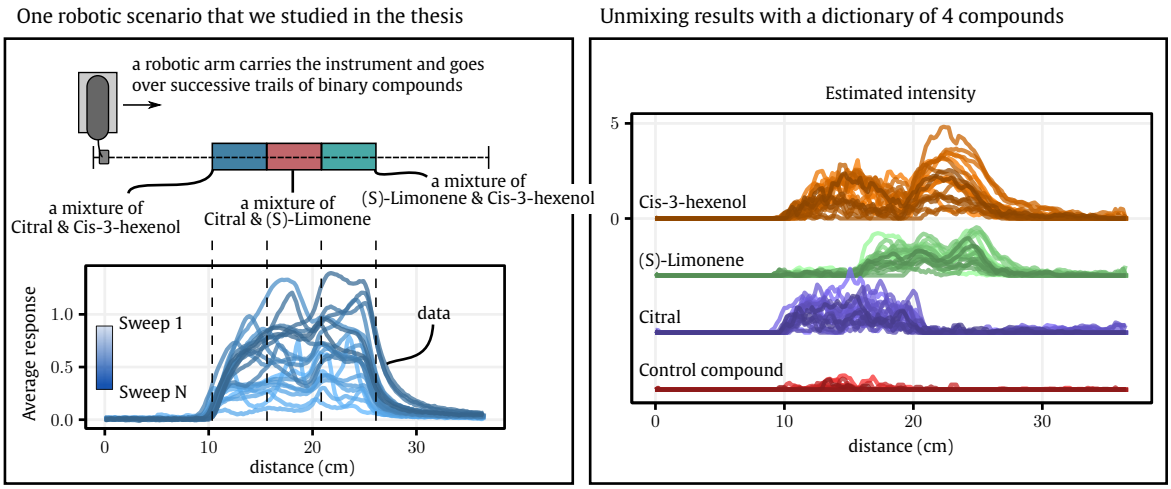


Fig. 7.3.: A robotic scenario in which gas mixtures occur. We placed three successive trails of binary compounds (*i.e.* composed of two different pure compounds) along a path. This path is followed by a robotic arm carrying the instrument. The result of a single sweep gives a single curve in left figure. In our experiments we carried out 20 sweeps which are represented by a color gradation (blue curves) in left figure. Right figure corresponds to the unmixing results (each line stands for a sweep) by using a dictionary of four compounds (three of them are actually present in the experiment while the fourth is not, we called it the control compound).

the task. When the full dictionary (12 different compounds) is considered, the algorithm performs poorly due to the existence of correlations between individual responses.

Future prospects

In all chapters, we discussed various perspectives of each work. In the following, we list the main future prospects and we split them in short and long term perspectives.

"Short-term" perspectives. In the near future, much effort must be made to develop an experimental setup enabling to reliably measure the concentration of volatile compounds. Ideally, this setup will also deal with gas mixtures. This improvement will allow to more precisely evaluate the performance of unmixing algorithms, and will especially allow to know if the estimated intensities in our work (see Figure 7.3) are directly related to the concentrations. By knowing the concentration, we will be also capable of studying the drift for gas quantification purpose. For quantification to be possible, we need calibration curves (sample points for which we know the concentration) which will be generated during the first measurement session. But how can we reuse these calibration curves for new sessions from which, for instance, we do not have any calibrating samples?

Finally, an interesting perspective of our robotic applications is the combination of Gas-Chromatography (GC) with the optoelectronic nose. By using this association, more complex mixtures than the binary and ternary cases can be easily studied with the current technology. From our robotic study, we can imagine that the analysis of mixtures of a dozen of different compounds is possible with the help of GC.

"Long-term" perspectives. The number of sensing materials (~ 20) currently used is by an order of magnitude lower than the number of different receptor types in the human nose (~ 400). Increasing this number is a long research work, implying the development of new materials which can be easily fixed over the surface, which show a good reversibility and which can be used for a long operation time. This research effort is fundamental to imagine a future instrument capable of what a human can do or even, of what an animal can do.

In Introduction, we mentioned the effect of the temperature on the instrument sensitivity. It would be interesting to investigate the temperature as a factor able to "virtually" increase the number of sensors. In this case, a single sensor would have several operating temperatures, say J, so one sensor would lead to J "virtual" sensors. Changing the operating temperature has in fact already been investigated with MOS-based systems, showing good results. Therefore, these results could be extended to Aryballe's technology. We actually know that the response is influenced by some chemical parameters which are themselves influenced by the temperature. By controlling this parameter, additional information about the chemical reactions may be created and then used for classification, quantification or Source Separation purposes.

Finally, our drift study requires also to be extended by considering several optoelectronic noses manufactured according to the same process. In this case, an inter-instrument drift will appear and will call for drift correction methods. In fact, in the future, the success of this kind of instruments relies on the possibility to use a single instrument for generating an entire data base used by other instruments in the field. This will pave the way for applications such as multi-robot exploration or in situ monitoring of pollutants using several measurement stations.

List of publications

Journal

[hal-02448737] P. Maho, C. Herrier, T. Livache, P. Comon, and S. Barthelmé
"*Real-time gas recognition and gas unmixing in robot applications*"
Submitted to: *Sensors and Actuators B: Chemical*
Under review n°2

[hal-02534216] P. Maho, C. Herrier, T. Livache, G. Rolland, P. Comon, and S. Barthelmé
"*Reliable chiral recognition with an optoelectronic nose*"
In: *Biosensors and Bioelectronics* 159 (Jul. 2020)
10.1016/j.bios.2020.112183

Conference

[hal-02180373] P. Maho, C. Dolcinotti, T. Livache, C. Herrier, A. Andreev, P. Comon, and S. Barthelmé. "*Reconnaissance de plusieurs composés chimiques à l'aide d'un robot équipé d'un nez électronique* ". In: *XXVIIème colloque Colloque francophone de traitement du signal et des images*, 2019, Lille (France)

[hal-02067954] P. Maho, C. Dolcinotti, T. Livache, C. Herrier, A. Andreev, P. Comon, and S. Barthelmé. "*Olfactive robot for gas discrimination over several months using a new optoelectronic nose* ". In: *18th International Symposium on Olfaction and Electronic Nose*, 2019, Fukuoka (Japan)
10.1109/ISOEN.2019.8823405

[hal-01802358] P. Maho, P. Comon, and S. Barthelmé. "*Non-linear source separation under the Langmuir model for chemical sensors*". In: *10th IEEE Workshop on Sensor Array and Multichannel Signal Processing*, 2018, Sheffield (United Kingdom)
10.1109/SAM.2018.8448636

Segmentation procedure for the data from robotic setups

A

In Chapters 4 & 6, we described several data sets obtained using a robotic setup. This setup is described in Chapter 2, Section 2.3.1. Briefly, a robot equipped with the Aryballe's eNose follows a path along which gas sources are disseminated. When the robot goes over a gas source, it generates a peak-shaped signal such the ones shown in Figure A.1. The location of the robot is not known over time and the robot does not interact with the eNose. Consequently, many temporal values are unknown, such as the beginning and the end of the injection when the robot goes over a gas source. These values are crucial for several algorithms that we developed during the thesis, so they need to be extracted. To that end, we introduce in this appendix a segmentation method.

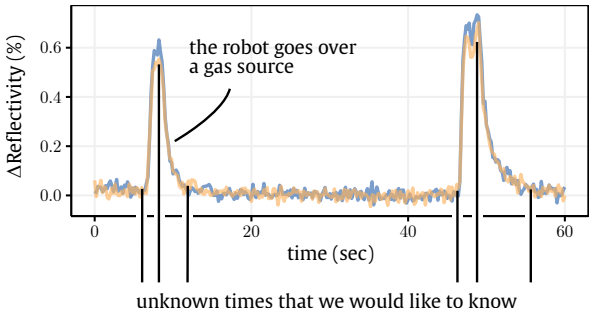


Fig. A.1.: Time series of two chemical sensors (blue and yellow curves) using the robotic setup described in Chapter 2, Section 2.3.1. In this figure, the robot went over two gas sources, corresponding to the two observed peaks.

We refer to segmentation as the extraction of Regions Of Interest (ROIs). A ROI is defined as the temporal range during which a VOC is picked up by the device. In the following, the segmentation is performed on the signal averaged across the chemical sensors. We assume that the baseline has been subtracted.

A.1 Synchronization of the laps

Recall that the data from Sniffer robot is the repetition of a predefined path, say that the robot did N laps. The gas sources are fixed cups so their positions do not change from one lap to another. Therefore, the whole signal can be seen as a pattern (one lap) which has been repeated N times.

To define a first approximation of the ROIs, we propose to work with the average lap, meaning the signal obtained by averaging the N signals of the N laps. This allows us to enhance the signal-to-noise ratio (SNR) and to make the detection of the ROIs easier. This first approximation is especially useful for the last laps where some gas sources are particularly difficult to detect due to VOC exhaustion (gas sources contain few drops of the compounds in liquid phase and this liquid solution may completely evaporate after several hours).

However, we mentioned earlier in this appendix that we do not know the location of the robot over time. Consequently, we need to first synchronize the laps, meaning that we need to identify the beginning of each lap.

A naïve solution could be to simply take as the beginning of each lap, a multiple of N_t (the duration of a lap). However, the robot course is not perfectly reproducible from one lap to another and the duration of a lap may vary a bit. In practice, we observed that this variation is small but small enough to correspond to a ROI ($\sim$1 second).

To have a better estimation of the beginning, we assume that we know the first lap, denoted by $y_{\text{ref}}(t)$ of duration N_t. In practice, this lap is manually extracted from the signals. Then, we define the sliding correlation $m(t) \in [-1, 1]$ as:

$$m(t) = \frac{\sum_{n=1}^{N_t} y_{\text{ref}}(n) y(n+t)}{(\sum_{n=1}^{N_t} y_{\text{ref}}(n)^2 \times \sum_{n=1}^{N_t} y(n+t)^2)^{\frac{1}{2}}} \tag{A.1}$$

$m(t)$ is then equal to 1 if there is a perfect match between the response $y(t)$ with the template $y_{\text{ref}}(t)$. From $m(t)$ we identify the beginning of each lap as the position of the maxima of $m(t)$, greater than a threshold ($= 0.7$).

A.2 Detection of the ROIs

The detection of the ROIs is a three-step procedure starting from the average lap. The first two steps correspond to a rough detection from the average lap based on a thresholding and on the derivative of the signal. The third step refines the ROI for each lap using a model.

First, from the average lap, we detect the maxima in a given temporal range (here, $\pm$5s due to the distance between two cups and due to the robot speed) which are greater than a threshold ($5\sigma_n$ with σ_n the standard deviation of the noise, estimated from the first points).

Second, from these maxima, we apply a heuristic procedure to approximate the ROIs. Starting from a local maximum, we increase the area around this position until there is a significant change in the sign of the derivative. This change signals the beginning of a new ROI. Thus, we identify $2R = 24$ ROIs which are represented by colored lines in Figure A.3.

Finally, we refine these ROIs using a model. We note $y_{rn}(t)$ the ROI of the average chemical sensor corresponding to the VOC r during the lap n at time t and $y_{prn}(t)$ the signal of the chemical sensor p.

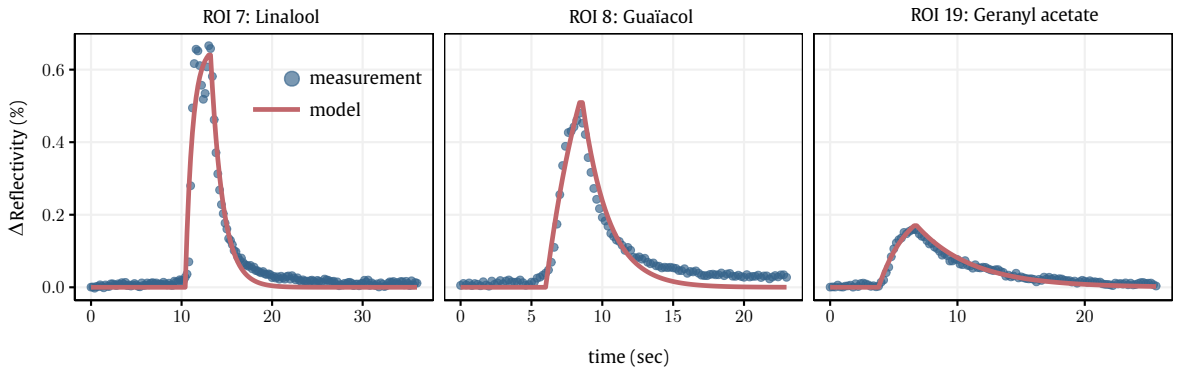


Fig. A.2.: Results of the fit for 3 different ROIs. The points correspond to the average chemical sensor and the lines correspond to the fit.

In Chapter 3, we saw that the response of a chemical sensor p to a VOC r can be fairly characterized by an exponential rise and decay (see Eq. (3.12) in Chapter 3). We use the same model but for modeling the average chemical sensor:

$$y_{rn}(t) = \begin{cases} 0 & \text{if } t < t^s_{rn} \\ \theta^{\text{eq}}_{rn}\left(1 - e^{-\tau^a_{rn}(t-t^s_{rn})}\right) = y_{a,rn}(t) & \text{if } t \in [t^s_{rn}, t^e_{rn}] \\ y_{a,rn}(t^e_{rn})e^{-\tau^d_{rn}(t-t^e_{rn})} & \text{if } t > t^e_{rn} \end{cases} \tag{A.2}$$

From model (A.2), we only want to identify the beginning t^s_{rn} and the end t^e_{rn} of the response of the VOC r for each lap n. This explains why we use the average chemical sensor: t^s_{rn} and t^e_{rn} are the same for each chemical sensor (they rise at the same time and decrease at the same time). However, compared to Chapter 3, the idea is not to use the model (A.2) for extracting features for classification (we should use each chemical sensor p for that).

For t^s_{rn}, we extract the adsorption signal by approximating t^e_{rn} as the time point with max amplitude. Then, for each possible value of t^s_{rn}, we estimate the parameters $(\theta^{\text{eq}}_{rn}, \tau^a_{rn})$ which minimize the quadratic cost, using the L-BFGS algorithm [Byr+95]. Finally, we choose t^s_{rn} as the one with the minimal cost.

For t^e_{rn}, the procedure is similar: assuming that t^s_{rn} is now known, we estimate the parameters $(\theta^{\text{eq}}_{rn}, \tau^a_{rn}, \tau^d_{rn})$ from the entire signal $y_{rn}(t)$ for values of t^e_{rn} close to the max, then we associate t^e_{rn} with the lowest cost. We this method, there is no need for an extra hyperparameter, such as a threshold value, for identifying the injection range in the ROI. In Figure A.2, we report some results of the fitting of the model for three different VOCs.

Finally, we refine each ROI by taking the points in $[t^s_{rn}, t^e_{rn} + \frac{1}{\tau^d_{rn}}]$. The final ROIs for one lap and one chemical sensor are represented by colored points in Figure A.3.

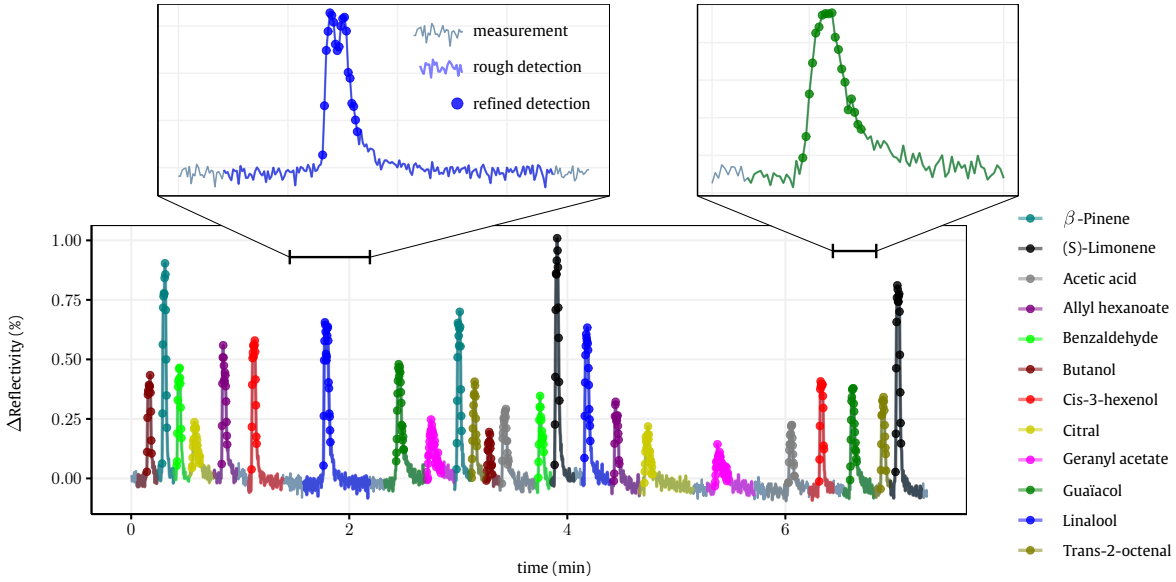


Fig. A.3.: Results of the segmentation for one lap and one chemical sensor. Each color stands for a different VOC (12 VOC in total). Colored lines correspond to rough first segmentation (based on the average lap) and colored points correspond to the final segmentation which is used in this thesis (after refining with the model (A.2)). Top figures correspond to zoom in on some parts of the response, left figure represents Linalool and right figure represents Guaïacol.

A.3 Outliers detection

Among all the ROIs extracted, some have too little signal to be useful, which is mainly due to VOC exhaustion. We detect and remove all the ROIs for which the mean of the average chemical sensor $y_{rn}(t)$ around the max (± 0.5 sec) is lower than $5\sigma_n$.

Identifiability of the Langmuir-SPRi model

B

B.1 Identifiability

In Chapter 3, we defined a theoretical model for describing the response of Aryballe's eNose. This model is the combination of two different models: a chemical part based on the Langmuir model and a physical part based on the linear approximation of the SPRi response. By noting $y(t)$ the response of a chemical sensor to a VOC, we recall the model:

$$y(t) = \begin{cases} 0 & \text{if } t < t_s \\ \nu \times \dfrac{ac}{d+ac}\left(1 - e^{-(d+ac)(t-t_s)}\right) = y_a(t) & \text{if } t \in [t_s, t_e] \\ y_a(t_e)e^{-d(t-t_e)} & \text{if } t > t_e \end{cases} \tag{B.1}$$

with ν a parameter related for instance to the VOC mass, a the adsorption constant, d the desorption constant and c the VOC concentration. t_s and t_e stand for the beginning of the VOC injection and for the beginning of the recovery phase. t_s and t_e are assumed to be known. We recall that all the parameters are non-negative. To lighten the notation, we removed the dependences of the parameters to the different dimensions of the problem (different chemical sensors p and different VOCs r). For the interested reader, these dependences appear in Eq. 3.10 (Chapter 3).

In this appendix, we study the identifiability of the model in Eq. B.1. In Section 3.1.4, we already discussed about the notion of identifiability and we highlighted a trivial indetermination due to the parameters a and c. We rewrite the model in Eq. B.1 by noting $z = ac$:

$$y(t, \boldsymbol{\zeta}) = \begin{cases} 0 & \text{if } t < t_s \\ \nu \times \dfrac{z}{d+z}\left(1 - e^{-(d+z)(t-t_s)}\right) = y_a(t, \boldsymbol{\zeta}) & \text{if } t \in [t_s, t_e] \\ y_a(t_e, \boldsymbol{\zeta})e^{-d(t-t_e)} & \text{if } t > t_e \end{cases} \tag{B.2}$$

$y(t, \boldsymbol{\zeta})$ means that the response over time is controlled by the set of parameters $\boldsymbol{\zeta} = (\nu, z, d)$. For the adsorption part, we found an additional indetermination. Let us define the following function:

$$T_\epsilon : \begin{bmatrix} \nu \\ z \\ d \end{bmatrix} \longmapsto \begin{bmatrix} \frac{\nu}{\epsilon} \\ z\epsilon \\ d + (1-\epsilon)z \end{bmatrix} \tag{B.3}$$

Then, for any $\epsilon \in \mathbb{R}^*$, we have $y_a(t, T_\epsilon(\boldsymbol{\zeta})) = y_a(t, \boldsymbol{\zeta})$. Solutions which are negative can be easily ruled out since the parameters are physical and chemical constants which are positive. However, there is still an infinite set of equivalent solutions.

To illustrate this indetermination, we can perform some numerical simulations. For that, we generate a response by considering only the adsorption part ($t_s = 0$ and we just use the second line of Eq. (B.2)) and using a vector $\boldsymbol{\zeta}^*$ (the ground truth). This response can be seen as a vector $\boldsymbol{y}_a \in \mathbb{R}^{N_t}$, which would have been measured after the interaction between a VOC and a chemical sensor. In practice, we do not know the vector of parameters $\boldsymbol{\zeta}^*$, so we need to estimate it from the "measured" response $\boldsymbol{y}_a$. For that, we need to define an optimization problem and solve it. To define an optimization problem, we need to choose a cost function (a function evaluating the error between the "measured" response and the response generated by the model with parameters $\boldsymbol{\zeta}$). Several cost functions can be used but the most classical one is the least squares cost function:

$$\hat{\boldsymbol{\zeta}} = \underset{\boldsymbol{\zeta} \geqslant 0}{\arg\min} \, \|\boldsymbol{y}_a - f_a(\boldsymbol{\zeta})\|_2^2 \tag{B.4}$$

where $\|\cdot\|_2^2$ is the ℓ_2-norm (square root of the sum of the squared terms) and $f_a(\boldsymbol{\zeta})$ is the adsorption function of parameters $\boldsymbol{\zeta}$ returning a vector of the same length than $\boldsymbol{y}_a$. The term $\boldsymbol{\zeta} \geqslant 0$ just means that we search for non-negative parameters.

There are a lot of different methods for solving Problem (B.4). For non-linear problems, the only possible choice is often to use iterative methods meaning that they start from an initial guess. From this initial guess, the algorithm finds another solution (*e.g.* by using derivatives) which is updated again and again until the algorithm converges (*e.g.* the cost evaluated on several successive iterations no longer changes). We use a quasi-Newton algorithm called L-BFGS [Byr+95] with the `R` package `optim`. A Newton algorithm is based on the use of the second derivatives and a quasi-Newton algorithm uses approximations to compute the second derivatives.

The algorithm needs then an initialization. This initialization is often random but a better initialization can also be estimated from the response. If we resort to random initial guesses, the non-identifiability of the model will cause a fundamental problem: the solution depends on the initialization. To illustrate this fact, we solve 2,000 times Problem (B.4) based on 2,000 random initializations. In Figure B.1, we represent the 2,000 initial guesses (hollow points) and the resulting solutions (yellow points) in the (ν, z) space and in the (d, z) space. The ground truth is represented by a red point. All the solutions are strictly equivalent due to the indetermination (the cost is null since we do not include any noise), so there is no way to tell which solution would be better. In Figure B.1, we also report two responses based on two possible $\boldsymbol{\zeta}$, one is the ground truth (red points) and the other is a possible solution (yellow line): different parameters give exactly the same response. Finally, we can notice that the set of possible solutions matches with the indetermination in Eq. (B.3) (black line).

This identifiability study reflects a major issue regarding the Langmuir-SPRi model: the adsorption part is not identifiable. However, it is interesting to note that the indetermination in Eq. (B.3) disappears if we consider both adsorption and desorption part. If we consider also desorption then the parameter d is identifiable by the exponential term (in the desorption

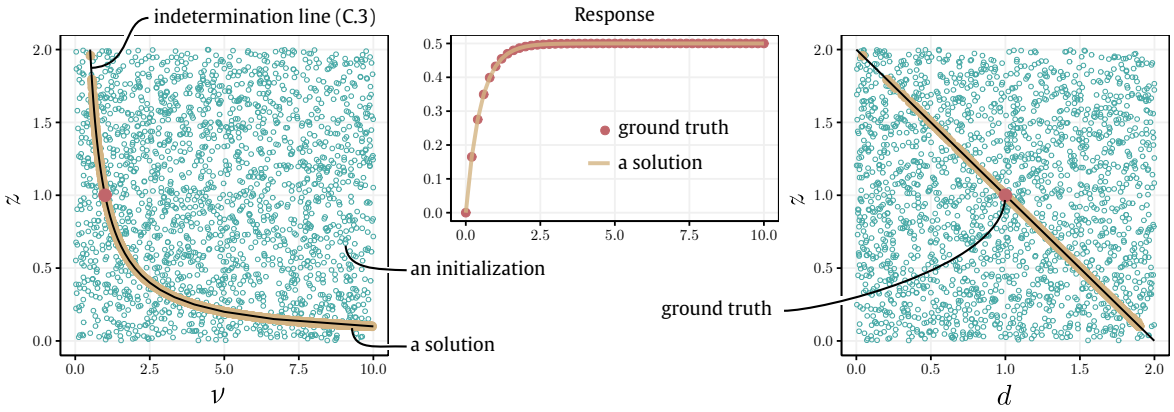


Fig. B.1.: Illustration of the non-identifiability of the adsorption model. Left and right figures: hollow points represent 2,000 random initializations and yellow points represent all the estimated solutions by solving Problem (B.4). The red point is the ground truth $\zeta^* = (1, 1, 1)$. Solutions and initializations are represented in the (ν, z) space (left) and in the (d, z) space (right). The black line is the indetermination in Eq. (B.3). Middle figure: we represent two responses generated from two different solutions, one is the ground truth (red points) and the other is a possible solution (yellow line).

part), meaning that it can no longer change and so the indetermination T_ϵ is only true for $\epsilon = 1$. To illustrate this point, we can redo the same numerical analysis than before. In this case, the response is generated by using both adsorption and desorption. A similar optimization problem than Problem (B.4) can be considered. In Figure B.2, we report the solutions based on 2,000 random initializations. It clearly shows that the indetermination T_ϵ (Eq. (B.3)) no longer exists.

However, there is still not a unique estimated solution, which is another problem. According to our simulations, these estimated solutions are local minima and these solutions are not strictly equivalent to the ground truth. In Figure B.2, we report two responses generated by two different solutions, one is the ground truth (red points) and the other is an estimated solution (yellow line). It clearly shows a data misfit when the response is generated using an estimated solution differing from the ground truth. In fact, out of 2,000 initializations, the algorithm converges $\sim$10% of the time to a local minimum. All these solutions had a non-null cost (but similar for all the local minima) showing that they are not strictly equivalent to the ground truth (which has a null cost) and do not correspond to a true indetermination in the full model. So, in this noise-free case, by running multiple times the algorithm from different initializations, it would be possible to correctly estimate the ground truth ζ^*, just by comparing the different costs. In the noisy case and on real data, these local minima can however become a real problem since the ground truth has no longer a null cost (due to the noise). It is especially true as all the models are just an approximation of what is occurring in real life. In other words, there is a "natural" data misfit which is not related to the noise but to all the approximations required for building the model. Consequently, even by considering the full model and multiple initializations, the cost may not be minimal for the true parameters and several solutions may give similar costs.

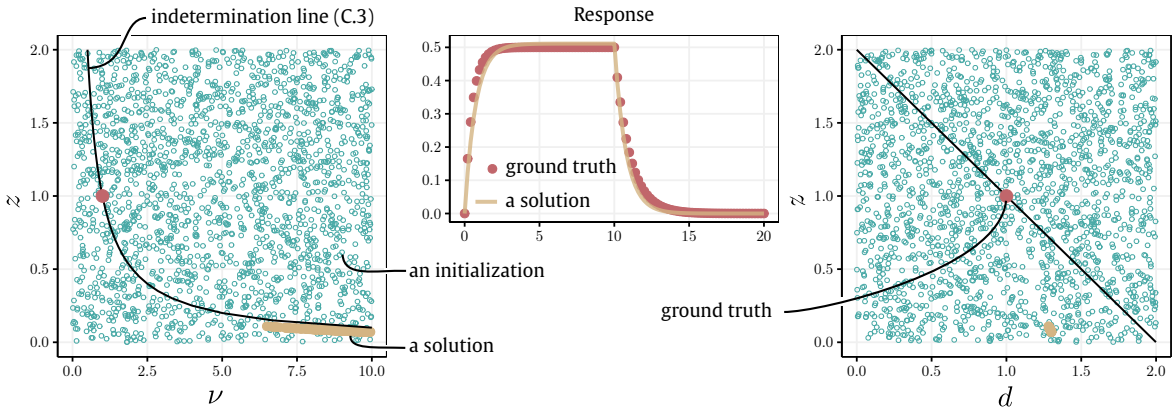


Fig. B.2.: Numerical simulations studying the identifiability of the full model. Left and right figures: hollow points represent 2,000 random initializations and yellow points represent all the estimated solutions (using a least-squares cost function and L-BFGS optimization method). The red point is the ground truth $\boldsymbol{\zeta}^* = (1, 1, 1)$. Solutions and initializations are represented in the (ν, z) space (left) and in the (d, z) space (right). The black line is the indetermination in Eq. (B.3). Middle figure: we represent two responses generated from two different solutions, one is the ground truth (red points) and the other is a possible solution (yellow line).

This study about the full model is not a thorough and theoretical proof of its identifiability, since the study relies on simulations (and a specific cost function). But, this forms an empirical piece of evidence that the full model is likely identifiable in the noise-free case. The existence of multiple local minima may however prevent estimation of the true parameters in the noisy case. In the next section, we describe a method enabling to improve the stability of the solution towards initialization.

B.2 A method to improve solution stability

By looking at the model in Eq. (B.2), we can notice an interesting property of the model: ν acts as a scaling factor. Consequently, by dividing by a quantity proportional to ν, we can remove the influence of this parameter from the response. To that end, we propose to create a new model by simply dividing by the integral of the adsorption part. The integration of the adsorption is straightforward:

$$\int_0^{t_e} y(t, \boldsymbol{\zeta}) \mathrm{d}t = \int_{t_s}^{t_e} y(t, \boldsymbol{\zeta}) \mathrm{d}t = \nu \times \frac{z}{d+z} \Big(t_e - t_s + \frac{1}{d+z} (e^{-(d+z)(t_e - t_s)} - 1) \Big) \tag{B.5}$$

By dividing the full model in Eq. (B.2) by this term, we can write a new model depending only on (z, d):

$$y(t, \tilde{\boldsymbol{\zeta}}) = \begin{cases} 0 & \text{if } t < t_s \\ \dfrac{1 - e^{-(d+z)(t - t_s)}}{t_e - t_s + \frac{1}{d+z}(e^{-(d+z)(t_e - t_s)} - 1)} = y_a(t, \tilde{\boldsymbol{\zeta}}) & \text{if } t \in [t_s, t_e] \\ y_a(t_e, \tilde{\boldsymbol{\zeta}}) e^{-d(t - t_e)} & \text{if } t > t_e \end{cases} \tag{B.6}$$

with $\tilde{\boldsymbol{\zeta}} = (z, d)$.

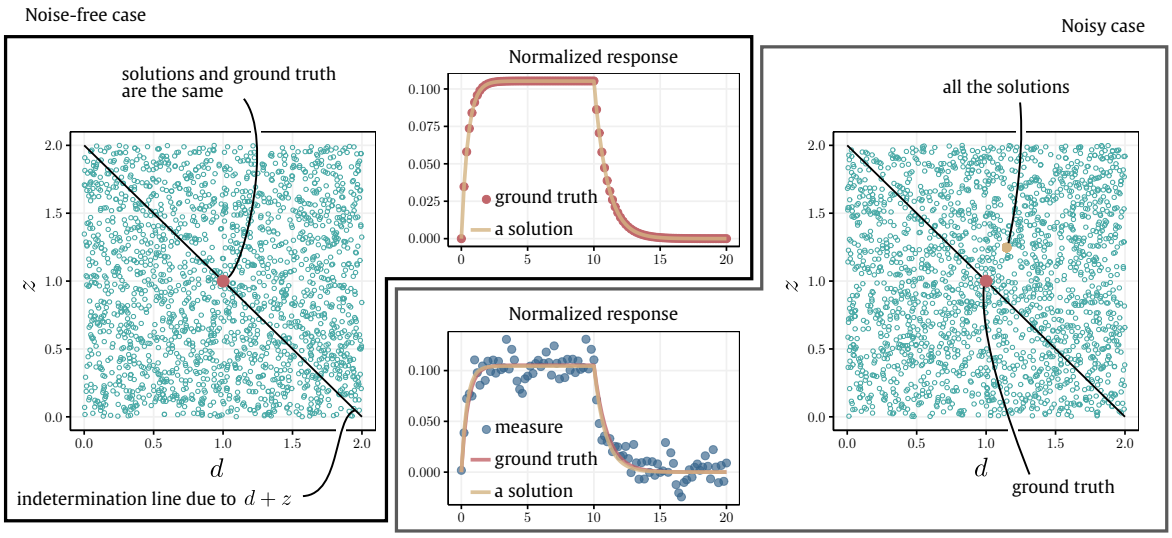


Fig. B.3.: Numerical simulations studying the identifiability of the full normalized model described in Eq. (B.6). The figure is divided in two parts: the noise-free case (left part) and the noisy case (right part). For both parts, hollow points represent 2,000 random initializations and yellow points represent all the estimated solutions (using a least-squares cost function and L-BFGS optimization method). The red point is the ground truth $\tilde{\boldsymbol{\zeta}}^* = (1, 1)$. The black line is the indetermination for the adsorption phase due to the term $d + z$. In the noise-free case, all the solutions converge to the ground truth (yellow points and the red point are superimposed). In the noisy case, all the solutions converge to the same solution, different from the ground truth due to the noise. Blue points in the middle figure represent the noisy measurement (additive Gaussian noise). The red and yellow lines correspond respectively to the noise-free response and to the response generated by the estimated parameters, showing little difference.

At first glance, the model in Eq. (B.6) seems much more complicated than the previous model in Eq. (B.2). However, this model is no longer dependent on three parameters but only two.

Despite this decrease in the number of parameters to estimate, the adsorption is still not identifiable. In fact, d and z always appear in a single term $d + z$. For any $\epsilon \in \mathbb{R}$, the two solutions (z, d) and $(z + \epsilon, d - \epsilon)$ are strictly equivalent.

However, if we again consider the full model (adsorption and desorption), this indetermination disappears for the same reason than in the previous section. d is identifiable thanks to the exponential term in the desorption. If d is identifiable then z becomes also identifiable. To empirically show it, we can redo the same simulations than in the previous section.

For that, we generate a response according to the full Langmuir-SPRi model, described in Eq. (B.2). This response is generated using a vector of parameters $\boldsymbol{\zeta}^* = (\nu^*, z^*, d^*)$. Then, we normalize this response by dividing by the integral of the adsorption part (computed from the response). The normalized response then follows the model in Eq. (B.6) described by the true parameters $\tilde{\boldsymbol{\zeta}}^* = (z^*, d^*)$. Using the normalized response, we estimate the parameters $\boldsymbol{\zeta}$ by solving a similar optimization problem than Problem (B.4).

In Figure B.3, we report the results of the simulations for both the noise-free case (left part) and the noisy case (right part). The noise is additive and Gaussian. In both cases, all the solutions estimated from 2,000 random initializations (hollow points) converge to the same

solution, indicating that the normalized model is much more stable than the previous model described in Eq. (B.2). In the noise-free case, all the solutions converge to the true parameters while in the noisy case the estimated solutions do not match with the ground truth due to the noise. As a final remark, once z and d have been estimated, we can also estimate ν by going back to the integral of the adsorption.

Robot olfaction - Additional results

C

C.1 Evaluation of the proposed constraints

In Chapter 6, we try to estimate a vector of intensities $\boldsymbol{c}(t) \in \mathbb{R}^R$, based on a measurement $\boldsymbol{y}(t) \in \mathbb{R}^P$ and a known dictionary $\boldsymbol{K} \in \mathbb{R}^{P \times R}$ (recall that, R: number of VOCs in the dictionary, P: number of chemical sensors). For that, we assume a linear model relating the response $\boldsymbol{y}(t)$ to the parameters $\boldsymbol{c}(t)$ and $\boldsymbol{K}$. To estimate $\boldsymbol{c}(t)$, we add some constraints to the optimization problem:

- Non-negativity (abbreviated Pos.): all the coefficients of $\boldsymbol{c}(t)$ are positive.
- Sparsity with ℓ_1-norm (abbreviated Sp.): many coefficients of $\boldsymbol{c}(t)$ are null.
- Support continuity (abbreviated Sup.): in a small time window (1 second in Chapter 6), we assume that the VOCs in the mixtures will slowly change. In other words, the position of the null coefficients will be "correlated" for all the vectors $\boldsymbol{c}(t)$ estimated in the time window.

In addition, a smoothing filter (average filter) is applied at the end of the estimation of $\boldsymbol{c}(t)$. We deem necessary to evaluate the benefits of each constraint and the benefits of the smoothing filter. For that, we evaluate 8 different methods which are listed in Table C.1.

Method	Constraints	Optimization problem	Smoothing filter?
L	-	$\arg\min_{\boldsymbol{c}} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2$	No
L + Smooth.	-	$\arg\min_{\boldsymbol{c}} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2$	Yes
L + Pos.	Non-negativity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2$	No
L + Pos. + Smooth.	Non-negativity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2$	Yes
L + Pos. + Sp.	Non-negativity Sparsity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$	No
L + Pos. + Sp. + Smooth.	Non-negativity Sparsity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$	Yes
L + Pos. + Sp. + Sup.	Non-negativity Sparsity Support continuity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$	No
L + Pos. + Sp. + Sup. + Smooth.	Non-negativity Sparsity Support continuity	$\arg\min_{\boldsymbol{c}\geqslant 0} \|\boldsymbol{y}(t) - \boldsymbol{K}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1$	Yes

Tab. C.1.: Methods compared in this section. For the abbreviations, L: Linear, Pos.: Non-negativity, Sp.: Sparsity, Sup.: Support continuity, Smooth.: Smoothing filter.

To compare these methods, we need some criterions. We propose 2 criterions: a classification score and the estimated number of VOCs present in the mixtures. The first criterion is detailed in Chapter 6 and tells us about a key aspect of the methods: their ability to correctly identify the gas sources. For comparing the methods which achieved the best classification scores, we can use another criterion which is important too: the estimated number of VOCs present in the mixtures. This criterion tells us about the ability of the method to correctly predict the number of VOCs actually present in the mixtures. In fact, it is an important property of our proposed method: we never assumed in the unmixing that we *a priori* know the actual number of VOCs being part of the mixtures. So this number is predicted by the algorithm and corresponds to the number of non-null intensities in $\boldsymbol{c}(t)$. To give only one value per method, we average the estimated number of VOCs when the robot is over a gas source.

Finally, we evaluate the methods on two data sets introduced in Chapter 6 and which can be considered as the most challenging data sets of the chapter:

1. The data set coming from Sniffer robot: 24 gas sources are disseminated in the environment. In this scenario, the closer to 1 the estimated number of VOCs, the better (because gas sources are pure and isolated) .
2. One data set coming from Sniffer arm: the robot goes over successive trails of binary compounds. In this scenario, the estimated number of VOCs must be between 2 and 3 (there are only binary compounds in the experiment but ternary mixtures can occur since we placed 3 different VOCs). For this data set, we will also evaluate the influence of the dictionary size.

C.1.1 Evaluation based on classification score

In Table C.2, we report the classification scores obtained by the different methods for the data from Sniffer robot. It clearly shows the benefit of using the smoothing filter for all the methods. It also shows a clear benefit of using the sparsity constraint. However, for this data set, there is no clear benefit of using the constraint of Support continuity when we apply a smoothing filter.

In Figure C.1, we report the classification scores obtained by the different methods for the data from Sniffer arm. In Chapter 6, we showed that the size of the dictionary is a crucial parameter regarding the unmixing performance. In Figure C.1 we therefore detail the scores for each possible dictionary size. It may be less easy to interpret due to the number of bars

Method	Classification score (%) Smoothing filter?	
	No	Yes
L	48.0	54.3
L + Pos.	61.8	65.2
L + Pos. + Sp.	70.9	72.7
L + Pos. + Sp. + Sup.	72.4	73.1

Tab. C.2.: Classification scores obtained by the different methods for the data from Sniffer robot.

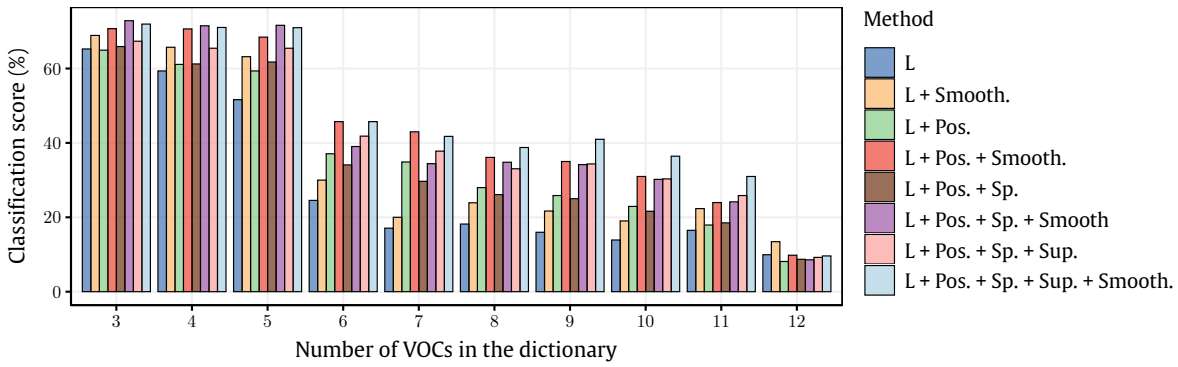


Fig. C.1.: Classification scores obtained by the different methods for the data from Sniffer arm (considering the scenario with successive trails of binary compounds). Each colored bar corresponds to one method. x-axis corresponds to the size of the dictionary: the greater the size, the harder the task.

represented in Figure C.1. However, the results obtained with Sniffer robot still hold in the case of Sniffer arm: the results with a smoothing filter perform significantly better whatever the dictionary size (for the full dictionary, meaning of size 12, all the methods perform poorly). In addition, these results allow us to draw another interesting conclusion: even if the benefit of the constraint of Support continuity is not clear for small dictionary sizes, the method with this constraint (in combination with the smoothing filter) performs significantly better when the size of the dictionary increases. Since in practice it can be unrealistic to consider only small dictionary size, it shows a clear benefit of our proposed method. Another point showing the interest of the method is that, the results with this method are always amongst the best classification scores.

To conclude, the criterion based on classification score indicates a clear advantage of the methods using a smoothing filter. Consequently, for the second criterion (based on the estimated number of VOCs), we consider only the methods using a smoothing filter. In addition, this study highlighted three methods which present good classification results: "L + Pos. + Smooth." (Linear model with non-negativity constraint followed by a smoothing filter), "L + Pos. + Sp. + Smooth." (Linear model with non-negativity and sparsity constraints followed by a smoothing filter) and "L + Pos. + Sp. + Sup. + Smooth." (Linear model with non-negativity, sparsity and support continuity constraints followed by a smoothing filter). So in the following, we focus on these three methods.

C.1.2 Evaluation based on the estimated number of VOCs

In Table C.3 and in Figure C.2, we respectively report the estimated number of VOCs for 3 methods for Sniffer robot and for Sniffer arm. In Table C.3, the closer to 1 the results, the better (because gas sources are pure and isolated). In Figure C.2, results between 2 and 3 are expected and the closer to 2 the results, the better (because gas sources are binary compounds).

Both Table C.3 and Figure C.2 indicate better results for the method that we proposed in Chapter 6. We can notice that for all the methods, the number of VOCs is overestimated when the dictionary becomes larger (Figure C.2).

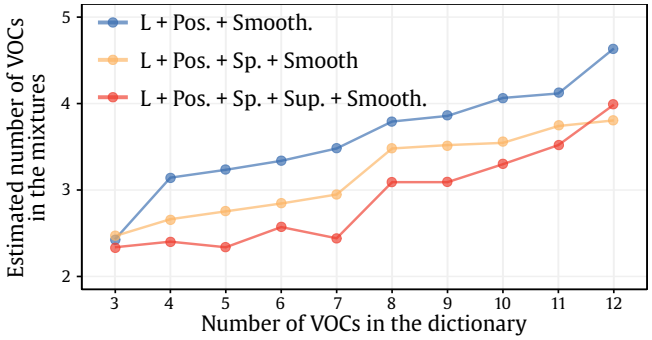


Fig. C.2.: Averaged estimated number of VOCs when the robot goes over a gas source for Sniffer arm. Recall that the estimated number must be between 2 and 3 (and it is likely that the closer to 2, the better). x-axis corresponds to the size of the dictionary: the greater the size, the harder the task.

Method	$\hat{N}_{\text{VOC}}$
L + Pos. + Smooth.	3.0
L + Pos. + Sp. + Smooth.	2.3
L + Pos. + Sp. + Sup. + Smooth.	1.7

Tab. C.3.: Averaged estimated number of VOCs when the robot goes over a gas source for Sniffer robot. Recall that the closer to 1 the estimated number of VOCs, the better.

Correctly estimate the number of VOCs in a mixture is not a trivial task and it is even less trivial given the conditions of our experiments. Consequently, we can conclude that the proposed method and all its constraints improve significantly the results that we would have obtained with less constrained methods.

C.2 Sniffer arm - Results for the sweep from right to left

	Citral		(S)-Limonene		Distance	
	Source 1	Source 2	Source 1	Source 2	Source 1	Source 2
Mean $\pm$ Stand. dev. (cm)	9.8 $\pm$ 0.32	29.1 $\pm$ 0.21	11.1 $\pm$ 0.21	24.9 $\pm$ 0.17	1.2 $\pm$ 0.28	4.2 $\pm$ 0.31
Ground-truth (cm)	[9,9.5]	[28,28.5]	[10.5,11]	[24.5,25]	1	3

Tab. C.4.: Estimation of the position of the isolated gas sources from Scenario ① with the default dictionary (Citral, (S)-Limonene, Guaïacol, Cis-3-hexenol), based on the location of the maximum intensity. The column distance refers to the spatial distance between Citral and (S)-Limonene. The two numbers of the ground truth correspond to the beginning and to the end of the scent strip (*i.e.* the gas source).

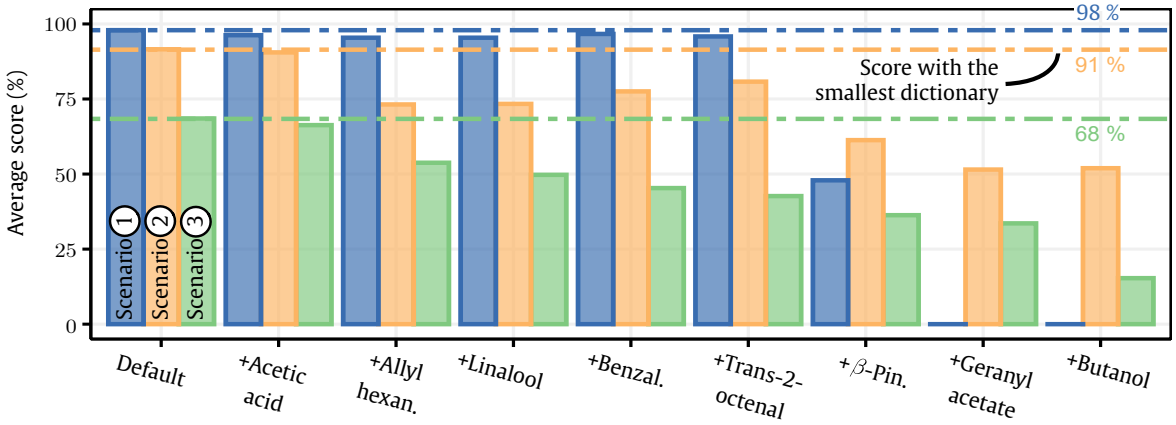


Fig. C.3.: Average score in each scenario and for each dictionary. A dashed line indicates the score obtained with the smallest dictionary (*e.g.* for Scenario ① the smallest dictionary is (Citral, (S)-Limonene)).

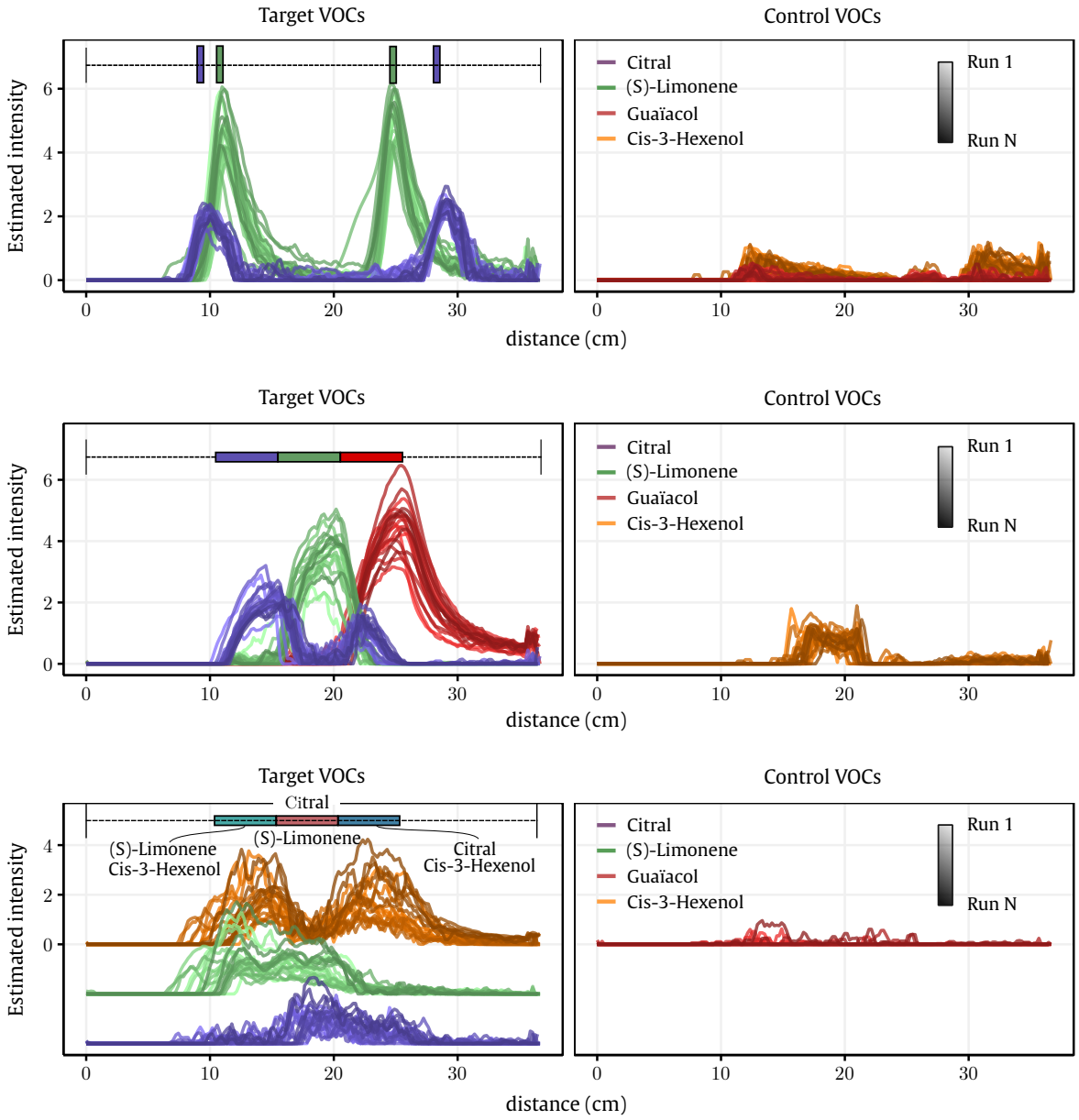


Fig. C.4.: Right to left direction. Results of the proposed algorithm for the different scenarios introduced. Each line corresponds to one sweep (first sweep is the lighter). Each color corresponds to the estimated intensity of the given VOC at the distance d. The results have been generated with the default dictionary ((S)-Limonene, Citral, Guaïacol, Cis-3-hexenol). For Scenario ③, the intensities of Citral and (S)-Limonene have been vertically shifted for easier visualization.

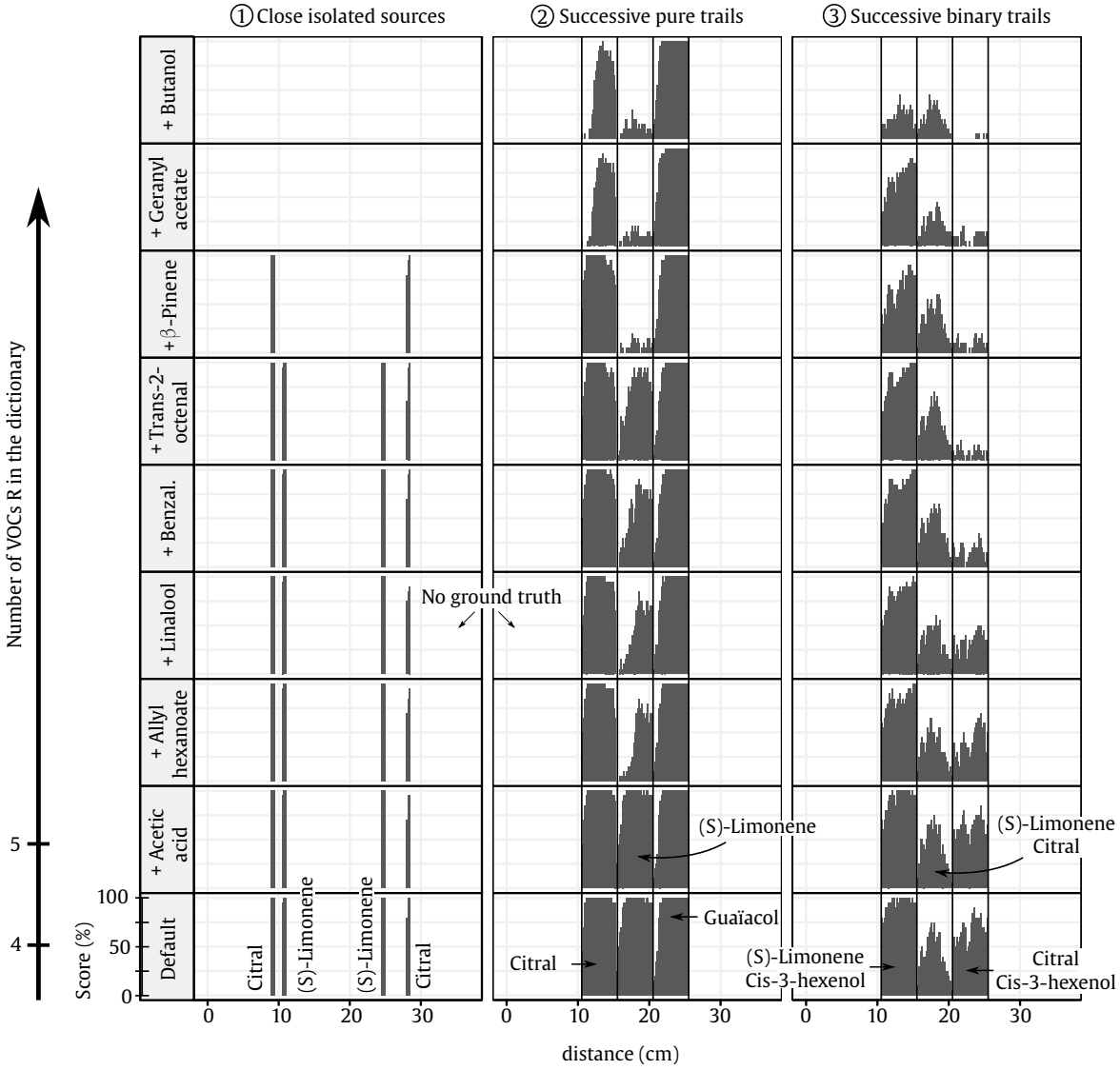


Fig. C.5.: Influence of the size of the dictionary on spatial distribution of the classification score. From the bottom up, the number of VOCs in the dictionary is increasing, starting from the default dictionary ((S)-Limonene, Citral, Guaïacol, Cis-3-hexenol). Left to right, each panel corresponds to the spatial distribution of the score for each scenario (for a position d, the score is the average across the 20 sweeps). The spatial distribution highlights the misclassifications (mainly in transition areas) in each scenario and for each dictionary. For **Scenarios** ② and ③, a black line indicates the theoretical transitions between one gas source to another.

Bibliography

[AC98] A. Agresti and B. A. Coull. “Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions”. In: *The American Statistician* 52.2 (May 1998), pp. 119–126 (cit. on p. 20).

[AJ05] S. Achard and C. Jutten. “Identifiability of post-nonlinear mixtures”. In: *IEEE Signal Processing Letters* 12.5 (May 2005), pp. 423–426 (cit. on p. 65).

[Alb+00] K. J. Albert, N. S. Lewis, C. L. Schauer, et al. “Cross-Reactive Chemical Sensor Arrays”. In: *Chemical Reviews* 100.7 (July 2000), pp. 2595–2626 (cit. on pp. 13, 18).

[And99] C. A. Andersson. “Direct orthogonalization”. In: *Chemometrics and Intelligent Laboratory Systems* 47.1 (Apr. 1999), pp. 51–63 (cit. on p. 84).

[Arc18] G. Archunan. “Odorant Binding Proteins: a key player in the sense of smell”. In: *Bioinformation* 14.1 (Jan. 2018), pp. 36–37 (cit. on p. 26).

[Ars+04] K. Arshak, E. Moore, G.M. Lyons, J. Harris, and S. Clifford. “A review of gas sensors employed in electronic nose applications”. In: *Sensor Review* 24.2 (Jan. 2004), pp. 181–198 (cit. on pp. 12, 13).

[Art+00] T. Artursson, T. Eklöv, I. Lundström, et al. “Drift correction for gas sensors using multivariate methods”. In: *Journal of Chemometrics* 14.5-6 (Sept. 2000), pp. 711–723 (cit. on pp. 79, 80, 82, 83, 98, 100, 104).

[AY05] B. W. Ache and J. M. Young. “Olfaction: Diverse Species, Conserved Principles”. In: *Neuron* 48.3 (Nov. 2005), pp. 417–430 (cit. on p. 3).

[Bas+03] N. Bassil, E. Maillart, M. Canva, et al. “One hundred spots parallel monitoring of DNA interactions by SPR imaging of polymer-functionalized surfaces applied to the detection of cystic fibrosis mutations”. In: *Sensors and Actuators B: Chemical* 94.3 (Oct. 2003), pp. 313–323 (cit. on p. 45).

[BD09] J. M. Bioucas-Dias. “A variable splitting augmented Lagrangian approach to linear spectral unmixing”. In: *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. Aug. 2009, pp. 1–4 (cit. on p. 140).

[Bec+10] R. Beccherelli, E. Zampetti, S. Pantalei, M. Bernabei, and K. C. Persaud. “Design of a very large chemical sensor system for mimicking biological olfaction”. In: *Sensors and Actuators B: Chemical* 146.2 (Apr. 2010), pp. 446–452 (cit. on p. 18).

[Bel57] Bellman, R. E. “Dynamic Programming”. In: *Princeton University Press* (1957) (cit. on p. 17).

[Bel61] R. E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Feb. 1961 (cit. on p. 17).

[Ben06] R. Bentley. “The Nose as a Stereochemist. Enantiomers and Odor”. In: *Chemical Reviews* 106.9 (Sept. 2006), pp. 4099–4112 (cit. on pp. 110, 111, 113, 126).

[Ber+12] M. Bernabei, K. C. Persaud, S. Pantalei, E. Zampetti, and R. Beccherelli. “Large-Scale Chemical Sensor Array Testing Biological Olfaction Concepts”. In: *IEEE Sensors Journal* 12.11 (Nov. 2012), pp. 3174–3183 (cit. on p. 16).

[Bey+99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. “When Is “Nearest Neighbor” Meaningful?” In: *Database Theory — ICDT’99*. Ed. by C. Beeri and P. Buneman. Lecture Notes in Computer Science. Springer, 1999, pp. 217–235 (cit. on p. 17).

[BHS97] C. M. Bishop, G. E. Hinton, and I. G. D. Strachan. “GTM through time”. In: *Artificial Neural Networks*. Jan. 1997, pp. 111–116 (cit. on p. 131).

[Bis06] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, 2006 (cit. on pp. 53, 90).

[Blo+15a] E. Block, S. Jang, H. Matsunami, V. S. Batista, and H. Zhuang. “Reply to Turin et al.: Vibrational theory of olfaction is implausible”. In: *Proceedings of the National Academy of Sciences* 112.25 (June 2015), E3155–E3155 (cit. on p. 2).

[Blo+15b] E. Block, S. Jang, H. Matsunami, et al. “Implausibility of the vibrational theory of olfaction”. In: *Proceedings of the National Academy of Sciences* 112.21 (May 2015), E2766–E2774 (cit. on p. 2).

[BOD12] K. Brudzewski, S. Osowski, and A. Dwulit. “Recognition of Coffee Using Differential Electronic Nose”. In: *IEEE Transactions on Instrumentation and Measurement* 61.6 (June 2012), pp. 1803–1810 (cit. on p. 16).

[Boe14] P. Boeker. “On ‘Electronic Nose’ methodology”. In: *Sensors and Actuators B: Chemical* 204.Supplement C (Dec. 2014), pp. 2–17 (cit. on pp. 8, 14, 16, 18, 21, 22).

[Bon+19] E. Bonah, X. Huang, J. H. Aheto, and R. Osae. “Application of electronic nose as a non-invasive technique for odor fingerprinting and detection of bacterial foodborne pathogens: a review”. In: *Journal of Food Science and Technology* (Nov. 2019) (cit. on p. 22).

[Bou+03] W. Bourgeois, A.-C. Romain, J. Nicolas, and R. M. Stuetz. “The use of sensor arrays for environmental monitoring: interests and limitations”. In: *Journal of Environmental Monitoring* 5.6 (2003), pp. 852–860 (cit. on p. 28).

[Bre18] S. Brenet. “Développement et optimisations d’un nez électronique basé sur l’imagerie de résonance de plasmons de surface”. PhD thesis. Université Grenoble Alpes, Oct. 2018 (cit. on pp. 25, 27, 129, 139).

[Bre+18] S. Brenet, A. John-Herpin, F.-X. Gallat, et al. “Highly-Selective Optoelectronic Nose Based on Surface Plasmon Resonance Imaging for Sensing Volatile Organic Compounds”. In: *Analytical Chemistry* (July 2018) (cit. on pp. 23, 26, 27, 133).

[Bro70] C. G. Broyden. “The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations”. In: *IMA Journal of Applied Mathematics* 6.1 (Mar. 1970), pp. 76–90 (cit. on p. 98).

[Bru+07] K. Brudzewski, J. Ulaczyk, S. Osowski, and T. Markiewicz. “Chiral behavior of TGS gas sensors: Discrimination of the enantiomers by the electronic nose”. In: *Sensors and Actuators B: Chemical* 122.2 (Mar. 2007), pp. 493–502 (cit. on pp. 21, 113, 114).

[BS10] V. E. Bochenkov and G. B. Sergeev. “Sensitivity, selectivity, and stability of gas-sensitive metal-oxide nanostructures”. In: *Metal Oxide Nanostructures and Their Applications*. Vol. 3. American Scientific Publishers, 2010, pp. 31–52 (cit. on p. 10).

[Byr+95] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. “A Limited Memory Algorithm for Bound Constrained Optimization”. In: *SIAM Journal on Scientific Computing* 16.5 (Sept. 1995), pp. 1190–1208 (cit. on pp. 49, 68, 142, 171, 174).

[CBC21] J. Cohen, R. Bro, and P. Comon. “Tensor decompositions in Food Sciences”. In: *Source Separation in Physical-Chemical Sensing*. Ed. by C. Jutten, L. T. Duarte, and S. Moussaoui. Wiley, 2021 (cit. on p. 82).

[Chy97] C. F. Chyba. “A left-handed Solar System?” In: *Nature* 389.6648 (Sept. 1997), pp. 234–235 (cit. on p. 111).

[CJ10] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic press, 2010 (cit. on p. 133).

[CLH12] R. P. Carpenter, D. H. Lyon, and T. A. Hasdell. *Guidelines for Sensory Analysis in Food Product Development and Quality Control*. Springer Science & Business Media, Dec. 2012 (cit. on p. 3).

[Com94] P. Comon. “Independent component analysis, a new concept?” In: *Signal processing* 36.3 (1994), pp. 287–314 (cit. on pp. 64, 65).

[CS96] J.-F. Cardoso and A. Souloumiac. “Jacobi Angles for Simultaneous Diagonalization”. In: *SIAM Journal on Matrix Analysis and Applications* 17.1 (Jan. 1996), pp. 161–164 (cit. on p. 83).

[CV95] C. Cortes and V. Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297 (cit. on p. 53).

[DA06] B. Domon and R. Aebersold. “Mass Spectrometry and Protein Analysis”. In: *Science* 312.5771 (Apr. 2006), pp. 212–217 (cit. on p. 4).

[DAS03] C. Distante, N. Ancona, and P. Siciliano. “Support vector machines for olfactory signals recognition”. In: *Sensors and Actuators B: Chemical* 88.1 (Jan. 2003), pp. 30–39 (cit. on p. 115).

[DC+11] S. Di Carlo, M. Falasconi, E. Sanchez, et al. “Increasing pattern recognition accuracy for chemical sensing by evolutionary based drift compensation”. In: *Pattern Recognition Letters* 32.13 (Oct. 2011), pp. 1594–1603 (cit. on pp. 79, 81, 84, 85, 87–90, 97, 98, 104).

[DCF12] S. Di Carlo and M. Falasconi. “Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges”. In: *Advances in Chemical Sensors*. InTech, 2012 (cit. on p. 77).

[DED06] C. M. Delahunty, G. Eyres, and J.-P. Dufour. “Gas chromatography-olfactometry”. In: *Journal of Separation Science* 29.14 (2006), pp. 2107–2125 (cit. on p. 6).

[DHS12] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Nov. 2012 (cit. on pp. 19, 20).

[Dic+01] M. E. Dickinson, G. Bearman, S. Tille, R. Lansford, and S. E. Fraser. “Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy”. In: *BioTechniques* 31.6 (Dec. 2001) (cit. on p. 140).

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38 (cit. on p. 90).

[DS04] D. Donoho and V. Stodden. “When does non-negative matrix factorization give a correct decomposition into parts?” In: *Advances in neural information processing systems*. 2004, pp. 1141–1148 (cit. on p. 73).

[DSP02] C. Distante, P. Sicilian, and K. C. Persaud. “Dynamic Cluster Recognition with Multiple Self-Organising Maps”. In: *Pattern Analysis & Applications* 5.3 (Aug. 2002), pp. 306–315 (cit. on p. 85).

[Fan+18] H. Fan, V. H. Bennetts, E. Schaffernicht, and A. J. Lilienthal. “A cluster analysis approach based on exploiting density peaks for gas discrimination with electronic noses in open environments”. In: *Sensors and Actuators B: Chemical* 259 (Apr. 2018), pp. 183–203 (cit. on p. 130).

[Fan+19] H. Fan, V. Hernandez Bennetts, E. Schaffernicht, and A. J. Lilienthal. “Towards Gas Discrimination and Mapping in Emergency Response Scenarios Using a Mobile Robot with an Electronic Nose”. In: *Sensors* 19.3 (Jan. 2019), p. 685 (cit. on pp. 129, 130, 137, 158).

[Fer+13] C. Ferdenzi, S. C. Roberts, A. Schirmer, et al. “Variability of Affective Responses to Odors: Culture, Gender, and Olfactory Knowledge”. In: *Chemical Senses* 38.2 (Feb. 2013), pp. 175–186 (cit. on p. 3).

[FHS18] X. Fu, K. Huang, and N. D. Sidiropoulos. “On Identifiability of Nonnegative Matrix Factorization”. In: *IEEE Signal Processing Letters* PP.99 (2018), pp. 1–1 (cit. on p. 73).

[FHT10] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of statistical software* 33.1 (2010), pp. 1–22 (cit. on p. 141).

[Fig13] Figaro. *TGS 2600 - for the detection of Air Contaminants*. 2013 (cit. on p. 133).

[Fis36] R. A. Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Annals of Eugenics* 7.2 (1936), pp. 179–188 (cit. on p. 53).

[Fle70] R. Fletcher. “A new approach to variable metric algorithms”. In: *The Computer Journal* 13.3 (Jan. 1970), pp. 317–322 (cit. on p. 98).

[Flu84] B. N. Flury. “Common Principal Components in k Groups”. In: *Journal of the American Statistical Association* 79.388 (Dec. 1984), pp. 892–898 (cit. on p. 83).

[FM71] L. Friedman and J. G. Miller. “Odor Incongruity and Chirality”. In: *Science* 172.3987 (June 1971), pp. 1044–1046 (cit. on pp. 111, 114).

[Fon+14] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, A. Vergara, and R. Huerta. “Chemical Discrimination in Turbulent Gas Mixtures with MOX Sensors Validated by Gas Chromatography-Mass Spectrometry”. In: *Sensors* 14.10 (Oct. 2014), pp. 19336–19353 (cit. on p. 131).

[Fon+15] J. Fonollosa, S. Sheik, R. Huerta, and S. Marco. “Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring”. In: *Sensors and Actuators B: Chemical* 215 (Aug. 2015), pp. 618–629 (cit. on pp. 16, 131, 132).

[Fon+16] J. Fonollosa, L. Fernández, A. Gutiérrez-Gálvez, R. Huerta, and S. Marco. “Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization”. In: *Sensors and Actuators B: Chemical* 236 (Nov. 2016), pp. 1044–1053 (cit. on p. 17).

[Gag+20] S. Gaggiotti, C. Hurot, J. S. Weerakkody, et al. “Development of an optoelectronic nose based on surface plasmon resonance imaging with peptide and hairpin DNA for sensing volatile organic compounds”. In: *Sensors and Actuators B: Chemical* 303 (Jan. 2020), p. 127188 (cit. on pp. 26, 27, 57).

[GB94] J. W. Gardner and P. N. Bartlett. “A brief history of electronic noses”. In: *Sensors and Actuators B: Chemical* 18.1 (Mar. 1994), pp. 210–211 (cit. on p. 7).

[GDR01] K. L. Goodner, J. G. Dreher, and R. L. Rouseff. “The dangers of creating false classifications due to noise in electronic nose and similar multivariate analyses”. In: *Sensors and Actuators B: Chemical* 80.3 (Dec. 2001), pp. 261–266 (cit. on p. 18).

[Gol70] D. Goldfarb. “A Family of Variable-Metric Methods Derived by Variational Means”. In: *Mathematics of Computation* 24.109 (1970), pp. 23–26 (cit. on p. 98).

[GT03] I. Gazit and J. Terkel. “Explosives detection by sniffer dogs following strenuous physical activity”. In: *Applied Animal Behaviour Science* 81.2 (Apr. 2003), pp. 149–161 (cit. on p. 3).

[Gue+00] P. Guedon, T. Livache, F. Martin, et al. “Characterization and Optimization of a Real-Time, Parallel, Label-Free, Polypyrrole-Based DNA Sensor by Surface Plasmon Resonance Imaging”. In: *Analytical Chemistry* 72.24 (Dec. 2000), pp. 6003–6009 (cit. on p. 23).

[HA04] M. Holmberg and T. Artursson. “Drift Compensation, Standards, and Calibration Methods”. In: *Handbook of Machine Olfaction*. Ed. by T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner. John Wiley & Sons, Ltd, 2004, pp. 325–346 (cit. on pp. 77, 78).

[HB+14] V. Hernandez Bennetts, E. Schaffernicht, V. Pomareda, et al. “Combining Non Selective Gas Sensors on a Mobile Robot for Identification and Mapping of Multiple Chemical Compounds”. In: *Sensors* 14.9 (Sept. 2014), pp. 17331–17352 (cit. on pp. 130, 131, 137, 158).

[HB16a] M. Hassan and A. Bermak. “Biologically Inspired Feature Rank Codes for Hardware Friendly Gas Identification With the Array of Gas Sensors”. In: *IEEE Sensors Journal* 16.14 (July 2016), pp. 5776–5784 (cit. on p. 17).

[HB16b] M. Hassan and A. Bermak. “Robust Bayesian Inference for Gas Identification in Electronic Nose Applications by Using Random Matrix Theory”. In: *IEEE Sensors Journal* 16.7 (Apr. 2016), pp. 2036–2045 (cit. on p. 17).

[HBZ06] A. Halperin, A. Buhot, and E. B. Zhulina. “On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions”. In: *Journal of Physics: Condensed Matter* 18.18 (2006), S463 (cit. on p. 43).

[HGO08] A. Hierlemann and R. Gutierrez-Osuna. “Higher-Order Chemical Sensing”. In: *Chemical Reviews* 108.2 (Feb. 2008), pp. 563–613 (cit. on pp. 10, 17, 18, 77, 78).

[HO96] N. Hansen and A. Ostermeier. “Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation”. In: *Proceedings of IEEE International Conference on Evolutionary Computation*. May 1996, pp. 312–317 (cit. on p. 84).

[Hol+97] M. Holmberg, F. A. M. Davide, C. Di Natale, et al. “Drift counteraction in odour recognition applications: lifelong calibration method”. In: *Sensors and Actuators B: Chemical* 42.3 (Aug. 1997), pp. 185–194 (cit. on pp. 16, 76).

[Hom06] J. Homola, ed. *Surface Plasmon Resonance Based Sensors*. Springer Series on Chemical Sensors and Biosensors. Springer-Verlag, 2006 (cit. on pp. 43, 46, 47, 60).

[Hou+12] Y. Hou, M. Genua, D. Tada Batista, et al. "Continuous Evolution Profiles for Electronic-Tongue-Based Analysis". In: *Angewandte Chemie International Edition* 51.41 (2012), pp. 10394–10398 (cit. on p. 23).

[HSS14] K. Huang, N. D. Sidiropoulos, and A. Swami. "Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition". In: *IEEE Transactions on Signal Processing* 62.1 (Jan. 2014), pp. 211–224 (cit. on pp. 64, 73).

[HTK00] J.-E. Haugen, O. Tomic, and K. Kvaal. "A calibration method for handling the temporal drift of solid state gas-sensors". In: *Analytica Chimica Acta* 407.1 (Feb. 2000), pp. 23–39 (cit. on pp. 77, 80).

[Hue+12] R. Huerta, S. Vembu, J. M. Amigó, T. Nowotny, and C. Elkan. "Inhibition in multiclass classification". In: *Neural Computation* 24.9 (Sept. 2012), pp. 2473–2507 (cit. on p. 131).

[Hug68] G. Hughes. "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1 (Jan. 1968), pp. 55–63 (cit. on p. 18).

[Hur+19] C. Hurot, S. Brenet, A. Buhot, et al. "Highly sensitive olfactory biosensors for the detection of volatile organic compounds by surface plasmon resonance imaging". In: *Biosensors and Bioelectronics* 123 (Jan. 2019), pp. 230–236 (cit. on pp. 26, 27, 57, 129).

[IWM12] H. Ishida, Y. Wada, and H. Matsukura. "Chemical Sensing in Robotic Applications: A Review". In: *IEEE Sensors Journal* 12.11 (Nov. 2012), pp. 3163–3173 (cit. on pp. 29, 128, 132).

[Jam+13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Vol. 6. Springer, 2013 (cit. on p. 148).

[JC82] A. K. Jain and B. Chandrasekaran. "39 Dimensionality and sample size considerations in pattern recognition practice". In: *Handbook of Statistics*. Vol. 2. Classification Pattern Recognition and Reduction of Dimensionality. Elsevier, Jan. 1982, pp. 835–855 (cit. on p. 21).

[JDM00] A. K. Jain, R. P. W. Duin, and J. Mao. "Statistical pattern recognition: a review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.1 (Jan. 2000), pp. 4–37 (cit. on p. 21).

[JHFT13] D. Johnen, W. Heuwieser, and C. Fischer-Tenhagen. "Canine scent detection—Fact or fiction?" In: *Applied Animal Behaviour Science* 148.3 (Oct. 2013), pp. 201–208 (cit. on p. 3).

[JHFT17] D. Johnen, W. Heuwieser, and Ca. Fischer-Tenhagen. "An approach to identify bias in scent detection dog testing". In: *Applied Animal Behaviour Science* 189 (Apr. 2017), pp. 1–12 (cit. on p. 3).

[Jia+17] S. Jiang, J. Wang, Y. Wang, and S. Cheng. "A novel framework for analyzing MOS E-nose data based on voting theory: Application to evaluate the internal quality of Chinese pecans". In: *Sensors and Actuators B: Chemical* 242 (Apr. 2017), pp. 511–521 (cit. on p. 17).

[JL09] I. M. Johnstone and A. Y. Lu. "On Consistency and Sparsity for Principal Components Analysis in High Dimensions". In: *Journal of the American Statistical Association* 104.486 (June 2009), pp. 682–693 (cit. on p. 18).

[JRP15] K. J. Johnson and S. L. Rose-Pehrsson. “Sensor Array Design for Complex Sensing Tasks”. In: *Annual Review of Analytical Chemistry* 8.1 (2015), pp. 287–310 (cit. on p. 18).

[Jun+98] L. S. Jung, C. T. Campbell, T. M. Chinowsky, M. N. Mar, and S. S. Yee. “Quantitative Interpretation of the Response of Surface Plasmon Resonance Sensors to Adsorbed Films”. In: *Langmuir* 14.19 (Sept. 1998), pp. 5636–5648 (cit. on p. 46).

[Khu+17] U. Khulal, J. Zhao, W. Hu, and Q. Chen. “Intelligent evaluation of total volatile basic nitrogen (TVB-N) content in chicken meat by an improved multiple level data fusion model”. In: *Sensors and Actuators B: Chemical* 238 (Jan. 2017), pp. 337–345 (cit. on p. 17).

[Koh82] T. Kohonen. “Self-organized formation of topologically correct feature maps”. In: *Biological Cybernetics* 43.1 (Jan. 1982), pp. 59–69 (cit. on p. 85).

[KP16] H. J. Ko and T. H. Park. “Bioelectronic nose and its application to smell visualization”. In: *Journal of Biological Engineering* 10.1 (Dec. 2016), p. 17 (cit. on p. 22).

[KS11] J. H. Kim and A. R. Scialli. “Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease”. In: *Toxicological Sciences* 122.1 (July 2011), pp. 1–6 (cit. on p. 111).

[Kyb+13] N. J. Kybert, M. B. Lerner, J. S. Yodh, G. Preti, and A. T. C. Johnson. “Differentiation of Complex Vapor Mixtures Using Versatile DNA–Carbon Nanotube Chemical Sensor Arrays”. In: *ACS Nano* 7.3 (Mar. 2013), pp. 2800–2807 (cit. on p. 113).

[Lak+14] S. Lakkis, R. Younes, Y. Alayli, and M. Sawan. “Review of recent trends in gas sensing technologies and their miniaturization potential”. In: *Sensor Review* 34.1 (Jan. 2014), pp. 24–35 (cit. on p. 13).

[Lan18] I. Langmuir. “The adsorption of gases on plane surfaces of glass, mica and platinum”. In: *Journal of the American Chemical Society* 40.9 (Sept. 1918), pp. 1361–1403 (cit. on pp. 42, 43).

[Lau+08] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. “Theorems on Positive Data: On the Uniqueness of NMF”. In: *Computational Intelligence and Neuroscience* (2008) (cit. on p. 73).

[LC06] A. Loutfi and S. Coradeschi. “Smell, think and act: A cognitive robot discriminating odours”. In: *Autonomous Robots* 20.3 (June 2006), pp. 239–249 (cit. on p. 130).

[LD04] A. Lilienthal and T. Duckett. “Building gas concentration gridmaps with a mobile robot”. In: *Robotics and Autonomous Systems*. European Conference on Mobile Robots (ECMR ’03) 48.1 (Aug. 2004), pp. 3–16 (cit. on pp. 130, 131).

[LGW17] Q. Li, Y. Gu, and N. Wang. “Application of Random Forest Classifier by Means of a QCM-Based E-Nose in the Identification of Chinese Liquor Flavors”. In: *IEEE Sensors Journal* 17.6 (Mar. 2017), pp. 1788–1794 (cit. on p. 17).

[Liv+15] T. Livache, A. Buhot, D. Bonnaffe, and Y. Hou-Broutin. “Electronic Nose or Tongue Sensors”. Pat. US20150037909A1. Feb. 2015 (cit. on p. 23).

[Llo+98] E. Llobet, X. Vilanova, J. Brezmes, et al. “Steady-State and Transient Behavior of Thick-Film Tin Oxide Sensors in the Presence of Gas Mixtures”. In: *Journal of The Electrochemical Society* 145.5 (Jan. 1998), pp. 1772–1779 (cit. on p. 157).

[Lou+05] A. Loutfi, M. Broxvall, S. Coradeschi, and L. Karlsson. “Object recognition: A new application for smelling robots”. In: *Robotics and Autonomous Systems* 52.4 (Sept. 2005), pp. 272–289 (cit. on p. 130).

[Lou+09] A. Loutfi, S. Coradeschi, A. J. Lilienthal, and J. Gonzalez. “Gas distribution mapping of multiple odour sources using a mobile robot”. In: *Robotica* 27.02 (2009), pp. 311–319 (cit. on p. 130).

[LT11] J. S. Lomond and A. Z. Tong. “Rapid Analysis of Dissolved Methane, Ethylene, Acetylene and Ethane using Partition Coefficients and Headspace-Gas Chromatography”. In: *Journal of Chromatographic Science* 49.6 (July 2011), pp. 469–475 (cit. on p. 159).

[Mab88] G. A. Mabbott. “An analogy for teaching interpretation of mass spectra”. In: *Journal of Chemical Education* 65.12 (Dec. 1988), p. 1052 (cit. on p. 5).

[Mad+18] S. Madrolle, L. T. Duarte, P. Grangeat, and C. Jutten. “A Bayesian Blind Source Separation Method for a Linear-quadratic Model”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Sept. 2018, pp. 1242–1246 (cit. on pp. 73, 158).

[Mai04] E. Maillart. “Imagerie par résonance des plasmons de surface pour l’analyse simultanée de multiples interactions biomoléculaires en temps réel”. PhD thesis. Université Paris Sud-Paris XI, 2004 (cit. on pp. 43, 45).

[Mal+99] B. Malnic, J. Hirono, T. Sato, and L. B. Buck. “Combinatorial Receptor Codes for Odors”. In: *Cell* 96.5 (Mar. 1999), pp. 713–723 (cit. on p. 2).

[Mar14] S. Marco. “The need for external validation in machine olfaction: emphasis on health-related applications”. In: *Analytical and Bioanalytical Chemistry* 406.16 (June 2014), pp. 3941–3956 (cit. on pp. 16, 18, 20, 27).

[Mar+98] S. Marco, A. Ortega, A. Pardo, and J. Samitier. “Gas identification with tin oxide sensor array and self-organizing maps: adaptive correction of sensor drifts”. In: *IEEE Transactions on Instrumentation and Measurement* 47.1 (Feb. 1998), pp. 316–321 (cit. on p. 85).

[MB31] E. C. Markham and A. F. Benton. “The adsorption of gas mixtures by silica”. In: *Journal of the American Chemical Society* 53.2 (Feb. 1931), pp. 497–507 (cit. on pp. 57, 59).

[MBC18] P. Maho, S. Barthelme, and P. Comon. “Non-linear source separation under the Langmuir model for chemical sensors”. In: *10th IEEE Workshop on Sensor Array and Multichannel Signal Processing (SAM 2018)*. Sheffield, United Kingdom: IEEE, July 2018 (cit. on pp. 64, 158).

[MBD01] G. Morrot, F. Brochet, and D. Dubourdieu. “The Color of Odors”. In: *Brain and Language* 79.2 (Nov. 2001), pp. 309–320 (cit. on p. 22).

[MBM19] D. Martinez, J. Burgués, and S. Marco. “Fast Measurements with MOX Sensors: A Least-Squares Approach to Blind Deconvolution”. In: *Sensors* 19.18 (Jan. 2019), p. 4029 (cit. on pp. 144, 158).

[Mea+13] M. I. Mead, O. A. M. Popoola, G. B. Stewart, et al. “The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks”. In: *Atmospheric Environment* 70 (May 2013), pp. 186–203 (cit. on pp. 129, 130).

[MGG12] S. Marco and A. Gutierrez-Galvez. “Signal and data processing for machine olfaction and chemical sensing: a review”. In: *IEEE Sensors Journal* 12.11 (2012), pp. 3189–3214 (cit. on pp. 79, 116, 133, 158).

[MGJ17] J. G. Monroy and J. Gonzalez-Jimenez. “Gas classification in motion: An experimental analysis”. In: *Sensors and Actuators B: Chemical* 240 (Mar. 2017), pp. 1205–1215 (cit. on pp. 17, 131).

[Mie96] P. Mielle. "'Electronic noses': Towards the objective instrumental characterization of food aroma". In: *Trends in Food Science & Technology* 7.12 (Dec. 1996), pp. 432–438 (cit. on pp. 16, 22).

[Mit97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997 (cit. on pp. 19, 20).

[MM10] E. Moser and M. McCulloch. "Canine scent detection of human cancers: A review of methods and accuracy". In: *Journal of Veterinary Behavior* 5.3 (May 2010), pp. 145–152 (cit. on p. 3).

[MML00] P. Mielle, F. Marquis, and C. Latrasse. "Electronic noses: specify or disappear". In: *Sensors and Actuators B: Chemical*. Proceedings of the International Symposium on Electronic Noses 69.3 (Oct. 2000), pp. 287–294 (cit. on pp. 16, 22).

[MMS19] H. M. McNair, J. M. Miller, and N. H. Snow. *Basic Gas Chromatography*. John Wiley & Sons, Sept. 2019 (cit. on p. 4).

[MNA02] L. Marques, U. Nunes, and A. T. de Almeida. "Olfaction-based mobile robot navigation". In: *Thin Solid Films*. Proceedings from the International School on Gas Sensors in conjunction with the 3rd European School of the NOSE Network 418.1 (Oct. 2002), pp. 51–58 (cit. on pp. 129, 131).

[MNM02] W. Maass, T. Natschläger, and H. Markram. "Real-time computing without stable states: a new framework for neural computation based on perturbations". In: *Neural Computation* 14.11 (Nov. 2002), pp. 2531–2560 (cit. on p. 132).

[Mon+16] J. G. Monroy, E. J. Palomo, E. López-Rubio, and J. Gonzalez-Jimenez. "Continuous chemical classification in uncontrolled environments with sliding windows". In: *Chemometrics and Intelligent Laboratory Systems* 158 (Nov. 2016), pp. 117–129 (cit. on p. 130).

[Mou+06] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. "Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling". In: *IEEE Transactions on Signal Processing* 54.11 (Nov. 2006), pp. 4133–4145 (cit. on pp. 73, 140).

[MP+08] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller. "Better than random? A closer look on BCI results". In: *International journal of bioelectromagnetism* 10.1 (2008), pp. 52–55 (cit. on p. 20).

[MT11] K. Mcgill and S. Taylor. "Robot Algorithms for Localization of Multiple Emission Sources". In: *ACM Comput. Surv.* 43.3 (Apr. 2011) (cit. on pp. 129, 130).

[MWD58] I. G. Mc William and R. A. Dewar. "Flame Ionization Detector for Gas Chromatography". In: *Nature* 181.4611 (Mar. 1958), pp. 760–760 (cit. on p. 7).

[NC+17] E. Núñez Carmona, V. Sberveglieri, A. Ponzoni, et al. "Detection of food and skin pathogen microbiota by means of an electronic nose based on metal oxide chemiresistors". In: *Sensors and Actuators B: Chemical* 238 (Jan. 2017), pp. 1224–1230 (cit. on p. 17).

[NDD95] C. D. Natale, F. A. M. Davide, and A. D'Amico. "A self-organizing system for pattern classification: time varying statistics and sensor drift effects". In: *Sensors and Actuators B: Chemical*. Eurosensors VIII 27.1 (June 1995), pp. 237–241 (cit. on p. 85).

[NHPH06] L. A. Nguyen, H. He, and C. Pham-Huy. "Chiral Drugs: An Overview". In: *International Journal of Biomedical Science : IJBS* 2.2 (June 2006), pp. 85–100 (cit. on p. 111).

[NS+17] J. Navarro-Sánchez, A. I. Argente-García, Y. Moliner-Martínez, et al. “Peptide Metal–Organic Frameworks for Enantioselective Separation of Chiral Drugs”. In: *Journal of the American Chemical Society* 139.12 (Mar. 2017), pp. 4294–4297 (cit. on p. 123).

[Pad+10] M. Padilla, A. Perera, I. Montoliu, et al. “Drift compensation of gas sensor array data by Orthogonal Signal Correction”. In: *Chemometrics and Intelligent Laboratory Systems* 100.1 (Jan. 2010), pp. 28–35 (cit. on pp. 28, 76, 79, 80, 84, 87, 104, 120).

[Pal+19] J. Palacín, D. Martínez, E. Clotet, et al. “Application of an Array of Metal-Oxide Semiconductor Gas Sensors in an Assistant Personal Robot for Early Gas Leak Detection”. In: *Sensors (Basel, Switzerland)* 19.9 (Apr. 2019) (cit. on p. 131).

[PCB10] M. D. Plumbley, A. Cichocki, and R. Bro. “Non-negative mixtures”. In: *Handbook of Blind Source Separation*. Ed. by P. Comon and C. Jutten. Academic Press, 2010 (cit. on p. 67).

[PD82] K. Persaud and G. Dodd. “Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose.” In: *Nature* 299.5881 (1982), pp. 352–355 (cit. on pp. 7, 11, 12).

[Pea+06] T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner. *Handbook of machine olfaction: electronic nose technology*. John Wiley & Sons, 2006 (cit. on pp. 3, 11–13, 18).

[Pen+19] E. O. Pentsak, D. B. Eremin, E. G. Gordeev, and V. P. Ananikov. “Phantom Reactivity in Organic and Catalytic Reactions as a Consequence of Microscale Destruction and Contamination-Trapping Effects of Magnetic Stir Bars”. In: *ACS Catalysis* 9.4 (Apr. 2019), pp. 3070–3081 (cit. on p. 119).

[Qui+12] P. Quignon, M. Rimbault, S. Robin, and F. Galibert. “Genetics of canine olfaction and receptor diversity”. In: *Mammalian Genome* 23.1 (Feb. 2012), pp. 132–143 (cit. on p. 3).

[RAE14] RAE Systems. *The PID Handbook: Theory and Applications of Direct-Reading Photoionization Detectors*. Honeywell, May 2014 (cit. on p. 7).

[RBW08] F. Röck, N. Barsan, and U. Weimar. “Electronic nose: current status and future trends”. In: *Chemical reviews* 108.2 (2008), pp. 705–725 (cit. on pp. 13, 18).

[RH71] G. F. Russell and J. I. Hills. “Odor Differences between Enantiomeric Isomers”. In: *Science* 172.3987 (June 1971), pp. 1043–1044 (cit. on pp. 111, 114).

[RN10] A. C. Romain and J. Nicolas. “Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview”. In: *Sensors and Actuators B: Chemical*. Selected Papers from the 13th International Symposium on Olfaction and Electronic Nose 146.2 (Apr. 2010), pp. 502–506 (cit. on p. 78).

[RP+18] R. Rodríguez-Pérez, R. Cortés, A. Guamán, et al. “Instrumental drift removal in GC-MS data for breath analysis: the short-term and long-term temporal validation of putative biomarkers for COPD”. In: *Journal of Breath Research* 12.3 (Mar. 2018), p. 036007 (cit. on p. 16).

[RS00] N. A. Rakow and K. S. Suslick. “A colorimetric sensor array for odour visualization”. In: *Nature* 406.6797 (Aug. 2000), pp. 710–713 (cit. on pp. 13, 23).

[Rud18] A. Rudnitskaya. “Calibration Update and Drift Correction for Electronic Noses and Tongues”. In: *Frontiers in Chemistry* 6 (Sept. 2018) (cit. on p. 79).

[Rus01] R. A. Russell. “Survey of Robotic Applications for Odor-Sensing Technology”. In: *The International Journal of Robotics Research* 20.2 (Feb. 2001), pp. 144–162 (cit. on p. 133).

[Sai+18] T. Saidi, O Zaim, M. Moufid, et al. “Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects”. In: *Sensors and Actuators B: Chemical* 257 (Mar. 2018), pp. 178–188 (cit. on p. 17).

[SBE98] E. Schaller, J. O. Bosset, and F. Escher. “‘Electronic Noses’ and Their Application to Food”. In: *LWT - Food Science and Technology* 31.4 (Apr. 1998), pp. 305–316 (cit. on p. 12).

[Sch02] V. Schurig. “Chiral separations using gas chromatography”. In: *TrAC Trends in Analytical Chemistry* 21.9 (Sept. 2002), pp. 647–661 (cit. on p. 112).

[Sch+16] F.-M. Schleif, B. Hammer, J. G. Monroy, et al. “Odor recognition in robotics applications by discriminative time-series modeling”. In: *Pattern Analysis and Applications* 19.1 (Feb. 2016), pp. 207–220 (cit. on p. 131).

[Sha70] D. F. Shanno. “Conditioning of Quasi-Newton Methods for Function Minimization”. In: *Mathematics of Computation* 24.111 (1970), pp. 647–656 (cit. on p. 98).

[SP02] J. R. Stetter and W. R. Penrose. “Understanding Chemical Sensors and Chemical Sensor Arrays (Electronic Noses): Past, Present, and Future”. In: *Sensors Update* 10.1 (2002), pp. 189–229 (cit. on pp. 18, 22).

[Spe15] C. Spence. “Multisensory Flavor Perception”. In: *Cell* 161.1 (Mar. 2015), pp. 24–35 (cit. on p. 22).

[Ste78] J. R. Stetter. “A surface chemical view of gas detection”. In: *Journal of Colloid and Interface Science* 65.3 (July 1978), pp. 432–443 (cit. on p. 12).

[Ste+91] E. Stenberg, B. Persson, H. Roos, and C. Urbaniczky. “Quantitative determination of surface concentration of protein with surface plasmon resonance using radiolabeled proteins”. In: *Journal of colloid and interface science* 143.2 (1991), pp. 513–526 (cit. on p. 46).

[Tib11] R. Tibshirani. “Regression shrinkage and selection via the lasso: a retrospective”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282 (cit. on p. 157).

[TJ99] A. Taleb and C. Jutten. “Source separation in post-nonlinear mixtures”. In: *IEEE Transactions on Signal Processing* 47.10 (Oct. 1999), pp. 2807–2820 (cit. on p. 65).

[TP15] M. P. Tiwari and A. Prasad. “Molecularly imprinted polymer based enantioselective sensing devices: A review”. In: *Analytica Chimica Acta* 853 (Jan. 2015), pp. 1–18 (cit. on pp. 112, 113).

[Tri10] M. Trincavelli. “Gas Discrimination for Mobile Robots”. PhD thesis. Örebro University, 2010 (cit. on pp. 28, 129).

[Tri+11] M. Trincavelli, A. Vergara, N. Rulkov, et al. “Optimizing the Operating Temperature for an array of MOX Sensors on an Open Sampling System”. In: *AIP Conference Proceedings*. Vol. 1362. Sept. 2011, pp. 225–227 (cit. on p. 129).

[Tro14] M. Trojanowicz. “Enantioselective electrochemical sensors and biosensors: A mini-review”. In: *Electrochemistry Communications* 38 (Jan. 2014), pp. 47–52 (cit. on pp. 28, 112, 114).

[Tur+15] L. Turin, S. Gane, D. Georganakis, K. Maniati, and E. M. C. Skoulakis. “Plausibility of the vibrational theory of olfaction”. In: *Proceedings of the National Academy of Sciences* 112.25 (June 2015), E3154–E3154 (cit. on p. 2).

[Var15] N. Vargesson. “Thalidomide-induced teratogenesis: History and mechanisms”. In: *Birth Defects Research Part C: Embryo Today: Reviews* 105.2 (2015), pp. 140–156 (cit. on p. 112).

[Ver+12] A. Vergara, S. Vembu, T. Ayhan, et al. “Chemical gas sensor drift compensation using classifier ensembles”. In: *Sensors and Actuators B: Chemical* 166-167 (May 2012), pp. 320–329 (cit. on pp. 28, 78).

[Ver+13] A. Vergara, J. Fonollosa, J. Mahiques, et al. “On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines”. In: *Sensors and Actuators B: Chemical* 185 (Aug. 2013), pp. 462–477 (cit. on pp. 131, 158).

[Was+19] T. Wasilewski, D. Migoń, J. Gębicki, and W. Kamysz. “Critical review of electronic nose and tongue instruments prospects in pharmaceutical analysis”. In: *Analytica Chimica Acta* 1077 (Oct. 2019), pp. 14–29 (cit. on p. 24).

[Wee+19] J. S. Weerakkody, C. Hurot, S. Brenet, et al. “Opto-electronic nose - temperature and VOC concentration effects on the equilibrium response”. In: *2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*. May 2019 (cit. on pp. 26, 27, 129).

[Wee+20] J. S. Weerakkody, S. Brenet, T. Livache, et al. “Optical Index Prism Sensitivity of Surface Plasmon Resonance Imaging in Gas Phase: Experiment versus Theory”. In: *The Journal of Physical Chemistry C* 124.6 (Feb. 2020), pp. 3756–3767 (cit. on pp. 26, 27, 129).

[WG] B. M. Wise and N. B. Gallagher. *http://www.eigenvector.com/MATLAB/OSC.html* (cit. on p. 84).

[WJS01] J. A. Westerhuis, S. de Jong, and A. K. Smilde. “Direct orthogonal signal correction”. In: *Chemometrics and Intelligent Laboratory Systems* 56.1 (Apr. 2001), pp. 13–25 (cit. on p. 84).

[Wol+98] S. Wold, H. Antti, F. Lindgren, and J. Öhman. “Orthogonal signal correction of near-infrared spectra”. In: *Chemometrics and Intelligent Laboratory Systems* 44.1 (Dec. 1998), pp. 175–185 (cit. on p. 84).

[WSE01] S. Wold, M. Sjöström, and L. Eriksson. “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (Oct. 2001), pp. 109–130 (cit. on p. 82).

[Yan+15] J. Yan, X. Guo, S. Duan, et al. “Electronic Nose Feature Extraction Methods: A Review”. In: *Sensors* 15.11 (2015), pp. 27804–27831 (cit. on pp. 9, 52, 116).

[Zam+05] S. Zampolli, I. Elmi, J. Stürmann, et al. “Selectivity enhancement of metal oxide gas sensors using a micromachined gas chromatographic column”. In: *Sensors and Actuators B: Chemical* 105.2 (Mar. 2005), pp. 400–406 (cit. on p. 160).

[Ziy+10] A. Ziyatdinov, S. Marco, A. Chaudry, et al. “Drift compensation of gas sensor array data by common principal component analysis”. In: *Sensors and Actuators B: Chemical*. Selected Papers from the 13th International Symposium on Olfaction and Electronic Nose 146.2 (Apr. 2010), pp. 460–465 (cit. on pp. 79–81, 83, 104).

[Zup+04] M. Zuppa, C. Distante, P. Siciliano, and K. C. Persaud. “Drift counteraction with multiple self-organising maps for an electronic nose”. In: *Sensors and Actuators B: Chemical* 98.2 (Mar. 2004), pp. 305–317 (cit. on p. 85).

Abstract

In Nature, olfaction is a key sense used by most species of animals for detecting, tracking and recognizing odors in the environment. An electronic nose is an instrument that takes inspiration from natural olfaction in order to detect volatile compounds. The main characteristic of this kind of instruments is that they use weakly-specific chemical sensors. This weak specificity allows the device to be sensitive to a broad range of volatile compounds, making it useful for a large range of applications. However, these instruments are still not a widespread technology. The small number of sensors used and the lack of repeatability of the instrument over time are some possible causes. In addition, the weak specificity of the sensors is sometimes a liability. For instance, in the case of gas mixtures, each compound contributes to the response of a chemical sensor according to its contribution to the mixture. In this thesis, we tackle several of these issues using a new instrument developed by Aryballe. Compared to other systems, this device boards a large number of sensors and this number can be easily increased. This electronic nose has already shown promising results in laboratory conditions. In the same vein, the thesis reveals the ability of the instrument to tell two mirror molecules apart. However, an electronic nose is not meant to be used only in the laboratory but must be useful in everyday conditions, just like its biological counterpart. To that end, we have developed different robotic setups, mimicking the search for multiple gas sources in the environment. They have allowed us to study recognition performance and gas unmixing in realistic conditions. In this context, new algorithms have been designed to classify and unmix signals in real-time. Finally, the thesis has also studied the repeatability of the instrument over 9 months. Correction methods have been proposed and allow the use of the instrument to be greatly extended.

Résumé

Dans la Nature, l'olfaction est un sens clé du règne animal permettant la détection, le suivi et la reconnaissance d'un grand nombre d'odeurs présentes dans l'environnement. Un nez électronique est un instrument s'inspirant des principes de base de l'olfaction naturelle afin d'identifier des composés volatiles. La caractéristique principale de ce type d'instruments est l'utilisation de capteurs chimiques peu spécifiques. Cette faible spécificité leur permet d'être sensible à un nombre élevé de composés volatiles, laissant entrevoir un large champ d'applications. Cependant, ces appareils de mesure restent assez peu répandus. Le faible nombre de capteurs utilisés et le manque de répétabilité de l'instrument au cours du temps en sont des causes possibles. La faible spécificité n'est pas non plus sans inconvénient. Dans le cas de mélanges gazeux par exemple, chacun des composés contribue à la réponse d'un capteur chimique en fonction de sa contribution au mélange. Dans cette thèse, nous étudions plusieurs de ces problématiques à l'aide d'un nouvel instrument développé par Aryballe. L'appareil a l'avantage d'embarquer un grand nombre de capteurs dont la gamme peut être facilement enrichie. Ce nez électronique a d'ores et déjà montré des résultats prometteurs en conditions de laboratoire. Dans la lignée de ces résultats, la thèse démontre la capacité de l'instrument à différencier des molécules miroir l'une de l'autre. Cependant, l'instrument n'est pas voué à une utilisation réservée au laboratoire mais doit surtout s'avérer utile en situation réelle, tout comme son équivalent biologique. A cette fin, la thèse a permis de mettre au point plusieurs plateformes robotiques, imitant la recherche d'un certain nombre de sources odorantes dans l'environnement. Les capacités de reconnaissance de l'appareil et la problématique des mélanges gazeux ont ainsi pu être étudiées dans ces conditions réalistes. Dans ce contexte, de nouveaux algorithmes ont été élaborés afin de pouvoir classifier et démélanger en temps-réel les signaux issus de l'instrument. Enfin, la thèse a également pu étudier la répétabilité de l'instrument sur un total de 9 mois et proposer des méthodes de correction permettant ainsi une utilisation prolongée de l'appareil.