POLITECNICO DI TORINO

Master's Degree in Mechatronic Engineering

Master's Degree Thesis

Peak Selection in Fermentation Monitoring Using HS-GC-IMS: A Machine Learning and Topological Data Analysis Approach

Supervisor
Prof. Massimo Violante

Candidate
Jiahao Xu
315113

December 2024

Acknowledgements

First of all, I want to thank my supervisor, Prof. Violante, for giving me the chance to work on this thesis and for his guidance throughout the process.

I also want to thank Luca and Alessandro for their advice and for teaching me to see things from new perspectives. I'll always remember the time we spent preparing presentations together.

Thank you to Gabriele, Irisa, and Ruize as well. Working on projects and studying for exams with you was a memorable part of this journey.

Finally, I want to thank my family. Even though they have never been to Italy, their support and encouragement have meant everything to me.

Thank you all!

Summary

Fermented foods have become increasingly popular in recent years, but fermentation processes are difficult to monitor because of their dynamic environment and the complexity of volatile organic compounds (VOCs). Traditional fermentation monitoring methods often neglect to detect VOCs, an important feature of the fermentation process. HS-GC-IMS (Headspace Gas Chromatography-Ion Mobility Spectrometry) provides a non-invasive and sensitive method for VOC detection, but generates large amounts of complex data that are difficult to analyse manually.

To overcome this challenge, this study firstly optimises the pre-processing steps such as region of interest (ROI) identification and threshold selection, then combines Persistent Homology and Variable Importance Projection (VIP) scores of PLS-DA, and finally succeeds in identifying the important peaks of VOCs generated during the fermentation process. This approach reduces manual intervention and improves the efficiency and accuracy of complex data processing.

The contribution of this thesis is the combination of HS-GC-IMS with ML and TDA, which provides a new way for intelligent fermentation monitoring with potential application in the fermented food industry.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research background

Studies over the years have shown that fermented foods have a role to play in good health, improving nutrition and reducing the risk of disease[10]. This is why fermented foods have become an essential part of our lives. Fermentation is a food processing technology that utilizes the growth and metabolic activity of microorganisms (e.g. lactic acid bacteria and brewer's yeast)for the transformation of complex organic compounds into simpler compounds[7], resulting in the production of a wide range of volatile organic compounds(VOCs) with distinctive flavours.

Traditionally, the monitoring of the fermentation process has relied on the measurement of pH and lactic acid content[7]. However, due to the dynamics of microorganisms and the diversity of fermentation environments (e.g., temperature, pH, oxygen concentration, etc.), these metrics do not provide a complete picture of microbial metabolic activity during fermentation, and thus do not allow for accurate, real-time control of the fermentation process. For this reason, the development of technologies that can detect microbial metabolites (especially VOCs) during fermentation has become a major focus of research.

1.2 Research objectives and significance

In recent years, Headspace Gas Chromatography-Ion Mobility Spectrometry (HS-GC-IMS) has been widely used for the detection of VOCs. It works by feeding the VOCs in the sample into the GC system through headspace injection, then the column separates these compounds and finally enters the IMS system for detection. It has the following advantages[9]:

- **High sensitivity:** it can detect VOCs at low concentrations, and can detect fermentation dynamics more accurately.
- **High efficiency:** it can detect the samples in a few minutes and real-time monitoring can be achieved.

- **Non-Destructive Analysis:** he technique is less destructive to the sample and allows continuous sampling without affecting the fermentation process.

With the appearance of new technologies often comes a number of problems. Firstly, due to the high sensitivity of the machine, HS-GC-IMS contains a large amount of compound information with each data acquisition. As the fermentation time progresses, the data detected at each time point is accumulated, resulting in a huge data set. It is difficult to process this huge amount of data efficiently with traditional methods, so that the data analysis process is slow and does not meet the requirement of rapid response in production. Secondly, the number of peaks in the data is too high due to the continuous production of new VOCs during the fermentation process. It is difficult for traditional methods to automatically identify peaks that are significant to the fermentation process. Especially when analysing fermentation processes in different media, critical peaks may be masked by others that are not significant. Finally, since most of the analysis of data is now done manually. The analyser needs to manually select region of interest (ROI) in the data. This is not only time-consuming and labour-intensive, but also susceptible to subjective factors. It may lead to inaccurate analyses.

In order to solve the above problems, this paper takes the following 2 points as the objectives of the study:

1. Improving the methods of data pre-processing to enhance the automation of data analysis.

   Data pre-processing is crucial in the analysis of VOCs, which directly affects the quality of the subsequent data and the reliability of the analysis results. Therefore, this study will aim to develop and improve existing pre-processing methods, such as drift time alignment, automatic threshold selection, baseline correction and automatically selecting ROI.

2. Combining machine learning (ML) and topological data analysis techniques (TDA) to analyse the dataset and select significant peaks (VOCs) from it.

   The changes of VOCs during fermentation reflect the dynamics of microbial metabolic activities. Therefore identification and extraction of representative peaks is crucial to understand the fermentation state. In this study, ML methods, specifically VIP (Variable Importance Projection) in Partial Least Squares Discriminant Analysis (PLS-DA), will be used to quantify the contribution of individual peaks of fermentation processification to the model. In addition, in order to compensate for the shortcomings of traditional data analysis methods, the study will use TDA, in particular Persistent Homology (PH), to deal with complex geometrical structures in high-dimensional data. PH can help us identify important features with persistence in the dataset and ensure that the selected peaks are significant.

Automation of the preprocessing process is important because this step reduces human intervention and subjective judgement. It can process a large amount of sample data in a short period of time and after processing the samples all meet a uniform standard. It facilitates the application of machine learning algorithms afterwards. At the same time, it improves the accuracy and repeatability of data analysis. Secondly, the combination

of PLS-DA and PH technology enables the precise identification of VOCs peaks in the huge data set that are closely related to the fermentation process. This step not only filters out unimportant VOCs, but also allows researchers to quantitatively assess the fermentation status from the information of these key peaks. For example, when a significant peak appears, it may be a VOC that is not favourable to fermentation. At this point, researchers can change the temperature, humidity or oxygen concentration in the fermentation environment to suppress the intensity of the peak.

In this study, by improving the automation and intelligence of the fermentation process monitoring, not only the quality and stability of the products were improved, but also the waste of raw materials was reduced. It provides some useful values for researchers in the field of fermentation for practical application.

1.3 Thesis structure

The aim of this thesis was to optimise the pre-processing of HS-GC-IMS data, and then to combine ML and TDA techniques to select and visualise significant peaks in the pre-processed data. The overall framework of the workflow is depicted in Figure 1.1.

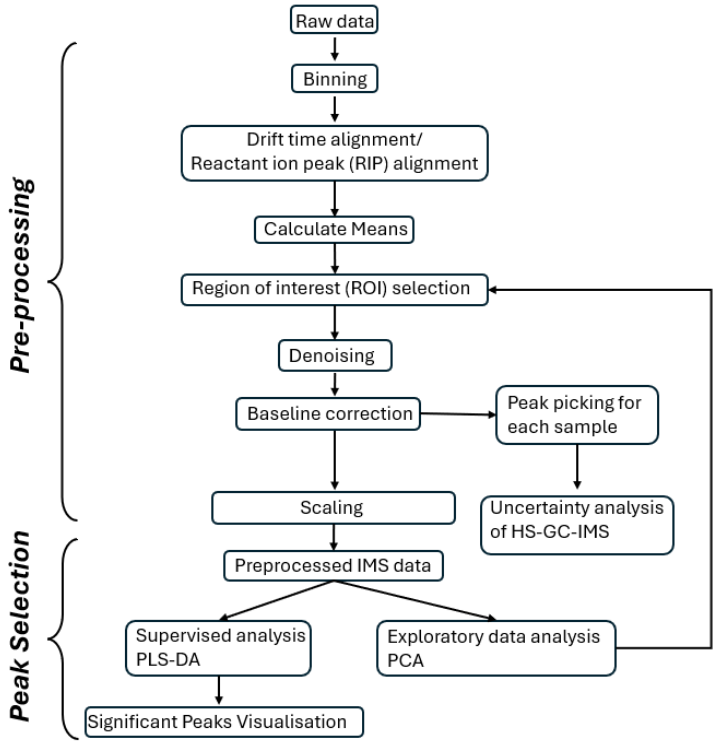


Figure 1.1: The workflow of HS-GC-IMS data analysis

We can observe that the workflow is divided into two main steps, pre-processing and peak selection. These two parts are also the parts that the paper focuses on explaining. The thesis is structured in the following chapters:

- **Chapter 1 Introduction**

  This chapter introduces the research background, objectives, and significance of this thesis. It also includes the overall structure of the thesis.

- **Chapter 2 Background and Literature**

  This chapter focuses on the principles of the HS-GC-IMS technique. The basic concepts of ML and PH are also introduced.

- **Chapter 3 Research Methodology**

  This chapter describes the experimental design and data processing methods in detail. Firstly, the medium preparation and sampling conditions are described. Secondly, the specific steps of pre-processing for HS-GC-IMS data analysis are discussed, including drift time alignment, automatic selection of regions of interest, automatic calculation of thresholds for denoising, and baseline correction. Finally, the method of peak selection(PH), and analysis tools such as PLS-DA and VIP score are introduced.

- **Chapter 4 Results and discussion**

  This chapter demonstrates firstly the results of the data pre-processing are shown and the results with the processing are analysed using PCA. Then the significant peaks in each sample were shown. The final part shows how the high scoring peaks in the VIP scores of the dataset changed during the fermentation process.

- **Chapter 5 Conclusion**

  This chapter summarises the main findings of this study and discusses the limitations of the study. Also, some recommendations are made for future research.

Chapter 2

Background and literature

2.1 HS-GC-IMS

HS-GC-IMS (Headspace Gas Chromatography-Ion Mobility Spectrometry) is an analytical technique mainly used for the detection of volatile organic compounds (VOCs).The technique combines three techniques, Headspace Sampling, Gas Chromatography (GC) and Ion Mobility Spectrometry (IMS), each of which performs a different function.

2.1.1 Headspace Sampling

Headspace sampling is a separation technique used to extract volatile compounds from the gas phase of a liquid or solid sample, as shown in Figure 2.1 These volatiles precipitate out of the sample under specific temperature and pressure conditions and accumulate in the headspace (the gas region above the sample) of the sample vessel and are subsequently injected into a gas chromatograph for analysis.[13]

Headspace sampling is performed as follows: first, a solid or liquid sample is placed in a headspace bottle and the bottle is sealed. The sample bottle is then heated in a heating device so that volatile compounds precipitate out of the sample and accumulate in the headspace region as the temperature increases. The bottle is then held for a certain period of time to allow the volatile compounds to reach equilibrium in the headspace. Finally, a headspace injector is used to extract the gas from the headspace and inject it into the gas chromatograph for analysis.

Headspace sampling analyses only the volatiles in the sample, avoiding the need to analyse non-volatile components in liquid or solid samples and therefore reducing column contamination or interference. Compared to other sample preparation techniques, headspace sampling usually does not require complex pre-treatment, reducing sample preparation time and reagent consumption. For all these reasons we can conclude that the advantages of this technique are fast, efficient and non-destructive analysis of volatiles, which is why this technique is widely used.

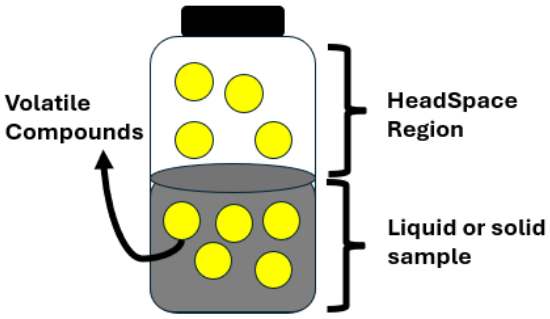


Figure 2.1: Diagram of headspace analysis

2.1.2 Gas Chromatography (GC)

Gas chromatography is a technique for separating mixtures. By injecting sample into a gas carrier (Mobile phase) and passing it through a column (Stationary phase) the chemicals in the mixture are separated. The retention time of each group of volatiles in the column is unique and we can use this to determine their identity.

The main components of a gas chromatograph are the injector, the carrier gas system, the column and Oven, the detector, as shown in Figure2.2.[18]

i. The injector is used to introduce the sample into the GC system, where it is then mixed with the carrier gas and entered into the system for separation and analysis. The injector must maintain consistent temperature and pressure to ensure that the sample is injected uniformly into the gas stream.

ii. The carrier gas is the mobile phase of gas chromatography that serves to carry the sample through the column. Carrier gases are usually helium, hydrogen or nitrogen because they are inert and do not react chemically with the sample. In addition, the purity of the carrier gas is very important to avoid introducing impurities that interfere with the analysis of the sample.

iii. A column is the heart of a gas chromatography system and is used to separate individual compounds in a sample. Its internal walls are coated with a stationary phase material, and the components of the sample are separated according to their different interaction behaviors with the stationary phase. Capillary columns are widely used for high-efficiency separations, and they are usually very thin and long, providing a high separation capacity.

iv. The column oven controls the temperature of the column. Temperature control is a key factor because the separation of different compounds is temperature dependent. With programmed temperature increase, the separation of different components in complex samples can be optimized to help improve resolution.

v. The detector is the final component of a gas chromatograph and is responsible for identifying and measuring the components passing through the column. Common

detectors are Flame Ionization Detectors (FID) and Thermal Conductivity Detectors (TCD) FID is very sensitive to organic compounds, while TCD is usually used to detect inorganic compounds or simple gases.

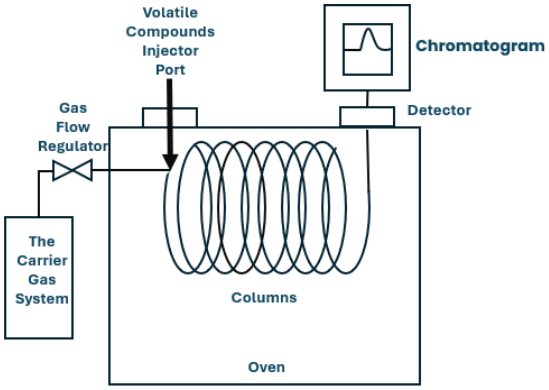


Figure 2.2: Simplified Structure of a Gas Chromatograph

The analysis process of the gas chromatograph is shown in Figure2.3. The mixture(shown in pink), made up of two components (A and B), is injected into the system through the injector port. This sample is carried by a gas through the chromatography column. Inside the column, the two components start to separate due to their different chemical properties, which affect how they interact with the stationary phase inside the column.

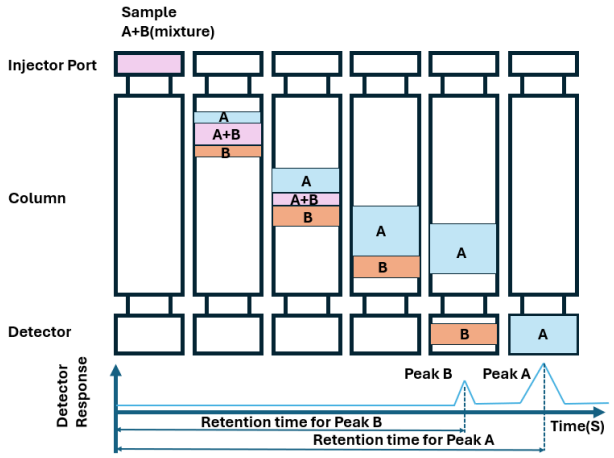


Figure 2.3: The analysis process of the gas chromatograph

As the components move through the column, component B (shown in orange) moves faster than component A (shown in blue), meaning it exits the column first. Once each component reaches the detector, it generates a signal, which is recorded as a peak on the chromatogram. The time it takes for each component to travel through the column and reach the detector is called the retention time.

In the lower part of the image, a graph shows the peaks for components A and B. Peak

B appears first, representing the faster eluting component, while Peak A appears later. The time difference between the peaks shows how the components were separated by the gas chromatography system. This method allows for the analysis and identification of the individual components in a mixture based on their retention times, as indicated by the formula (2.1) where t_m is the residence time of the compound in the mobile phase and t_s is the residence time of the compound in the stationary phase.

$$t_r = t_m + t_s \tag{2.1}$$

2.1.3 Ion Mobility Spectrometry (IMS)

Ion Mobility Spectrometry (IMS) is an analytical technique for the characterisation of chemical ionic compounds based on the difference in migration rates of different gas-phase ions in an electric field[19]. The mobility of an ion K is defined as the drift velocity v_d of the ion per unit electric field strength E, and is given by:

$$K = \frac{v_d}{E} \tag{2.2}$$

where:

- K is the ion mobility (in $\mathrm{cm}^2/\mathrm{V} \cdot \mathrm{s}$),
- v_d is the drift velocity (in cm/s),
- E is the electric field strength (in V/cm).

The time it takes for an ion to travel through the drift tube, known as the drift time t_d, can be expressed as:

$$t_d = \frac{L}{K \cdot E} \tag{2.3}$$

where L is the length of the drift tube.

The ion mobility (K) depends on the gas density (N) and temperature (T)[6], and is often expressed in reduced form:

$$K_0 = K \cdot \frac{N}{N_0} = K \cdot \left(\frac{p}{p_0}\right)\left(\frac{T_0}{T}\right) \tag{2.4}$$

where:

- K_0 is the reduced mobility (standardized at 760 Torr and 273.16 K),
- N_0 is the standard gas density,
- p_0 and T_0 are the standard pressure and temperature.

Another key aspect of IMS is the collision cross section (CCS), denoted as Ω, which provides structural information about the ion[6]. It is linked to ion mobility through the equation:

$$K = \frac{3ze}{16N}\sqrt{\frac{2\pi}{\mu k_B T}} \cdot \frac{1}{\Omega} \tag{2.5}$$

where:

- z is the ion's charge state,
- e is the elementary charge,
- N is the gas number density,
- μ is the reduced mass of the ion-gas pair,

$$\mu = \frac{mM}{m+M} \tag{2.6}$$

  where m is ion mass (analyte) and M is molecular mass(drift gas).

- k_B is the Boltzmann constant,
- T is the gas temperature.

Although the above formulas seem to be very complicated, in fact, in our operation, the length of the drift tube is fixed, and the flow rate and temperature of the drift gas are fixed, the ion mobility is only related to its own ion mass m and the CCS (Ω) (which is a constant), and this is the basis for the construction of ion mobility spectra.

Then we can understand the specific working principle of ion mobility spectroscopy through the following figure2.4. The drift tube is divided into two main regions, an ionization and reaction region and a drift region. The sample molecule first enters the system through the gas inlet. They are transformed into ions in the presence of β rays as they pass through the ionization source. From the ionization and reaction region, a constant electric field is applied in the IMS, which is used to drive the ions through the device. The ions then enter the drift region through the periodically opening Ion shutter. In this region, the ions move towards the Faraday plate on the right side under the influence of the electric field. The drift region is filled with drift gas(Nitrogen), and this gas provides resistance that causes the particles to move at different speeds in the drift region. The drift speed of the ions depends on their mass, shape and charge. Therefore, lighter or smaller ions will usually reach the end of the drift region faster than heavier and larger ions. Since the ionisation process is not 100% successful, there will always be some un-ionised molecules that enter the drift zone and interfere with the detection, and the drift gas will blow these molecules out of the gas outlet. In the drift tube we can see that there are several drift rings, which help to maintain the uniformity of the electric field and keep the molecules and ions in regular motion throughout the drift. The detector, usually a Faraday plate, is at the end of the drift zone. The Faraday plate is able to detect ions that reach the end point and generate a signal. The strength

of the signal is related to the number of ions, while the time it takes for the ions to reach the Faraday plate (drift time t_d) depends on the mobility of the ions K[12].

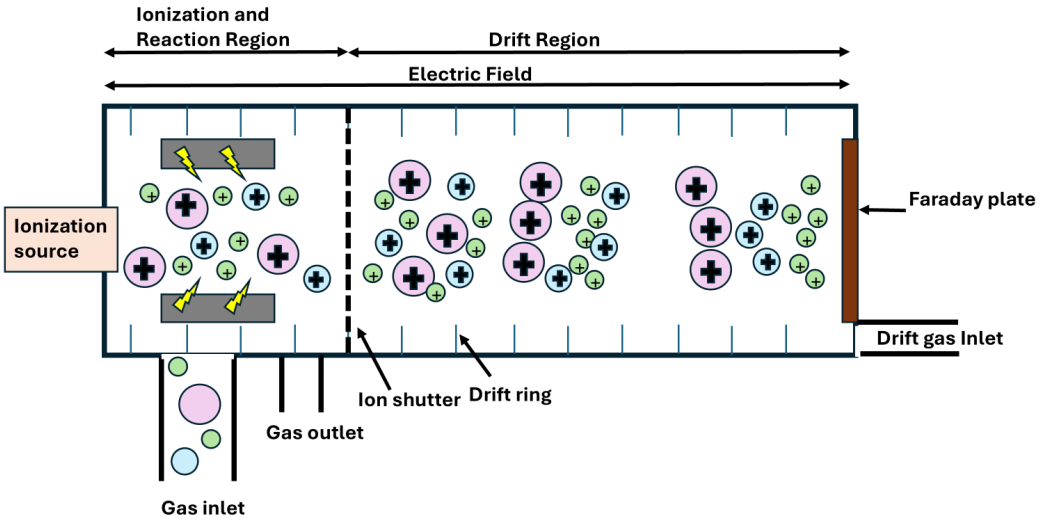


Figure 2.4: Ion Mobility Spectrometry (IMS) Schematic working principle

In IMS, the drift gas (nitrogen) is ionized in the ionization and reaction region at atmospheric pressure to produce a positively charged nitrogen ion and release electrons(equation2.7). After a series of reactions(from equation 2.8 to 2.12) these electrons is transferred to a trace amount of water, which is contained in the nitrogen itself, and finally a hydrated hydrogen ion($H^+(H_2O)_3$) is formed. These hydrated hydrogen ions continue to accumulate in the drift zone and produce a stable peak on the IMS detector - the reactive ion peak (RIP).

Formation of Reactant Ions - Hydrated Proton[3]

$$N_2 + e^- \longrightarrow {N_2}^+ + 2\,e^- \tag{2.7}$$
$${N_2}^+ + 2\,N_2 \longrightarrow {N_4}^+ + N_2 \tag{2.8}$$
$${N_4}^+ + H_2O \longrightarrow 2\,N_2 + {H_2O}^+ \tag{2.9}$$
$${H_2O}^+ + H_2O \longrightarrow H^+(H_2O) + OH \tag{2.10}$$
$$H^+(H_2O) + H_2O + N_2 \longrightarrow H^+(H_2O)_2 + N_2 \tag{2.11}$$
$$H^+(H_2O)_2 + H_2O + N_2 \longrightarrow H^+(H_2O)_3 + N_2 \tag{2.12}$$

The RIP peak is the background peak in the IMS and is used for calibration and reference. It provides a stable signal that can be used as a reference point in the absence of sample analytes. By observing the position and intensity of the RIP peak, you can ensure the consistency and accuracy of the IMS instrument under different conditions.

Combining the advantages of HS, GC and IMS technologies has made HS-GC-IMS one of the most popular analytical tools for the detection of volatile organic compounds in recent years. The advantages of this application are high sensitivity, non-destructive detection, fast response and the ability to handle complex sample matrices, which makes it particularly suitable for monitoring fermentation processes. HS-GC-IMS accurately captures the diverse and dynamically complex changes in VOCs produced during fermentation, helping to analyse key compounds and quality changes in the fermentation process, and contributing to the optimisation and control of the fermentation process.

The latest research in the analysis of GC-IMS data is the development of a python package gc-ims-tool by Joscha Christmann [2]. It provides reading of .mea files, preprocessing methods (drift time alignment, baseline alignment, etc.) and visualisation. However, it still uses manual parameter input for region of interest selection and threshold selection, which will make the subsequent analysis affected by subjective judgement. This thesis optimises the pre-processing process according to this problem by automating the ROI and threshold selection process. This reduces human intervention and allows all samples to be under a uniform standard.

2.2 Machine Learning

Machine learning is a subject that uses algorithms to automatically learn regularities from data and make predictions or classifications. Its algorithms can usually be divided into two main categories: supervised learning and unsupervised learning. Supervised learning uses labelled data for training to predict labels for new data and to discover potential relationships in labelled data; unsupervised learning is used to discover potential structures or patterns in unlabelled data. In food chemistry and fermentation research, high-dimensional data such as HS-GC-IMS often have complex multivariate features from which traditional analytical methods have difficulty extracting useful features, and machine learning can help to analyse such complex data. Through unsupervised learning, samples can be analysed exploratively to explore the internal structure of the data; through supervised learning, different types of fermentation samples can be classified and important features can be extracted, which in turn can explain the changes of chemicals during the fermentation process. Thus, machine learning helps to reveal in depth the chemical changes in the food fermentation process.

2.2.1 Unsupervised learning

Unsupervised learning is an analysis method that does not rely on labelled data and can help to discover underlying patterns and structures from data. One important part of unsupervised learning is Exploratory Data Analysis (EDA).At its core, EDA is the exploration of data without excessive preconceived assumptions or a priori knowledge to better understand patterns in the data, detect outliers or systematic errors, and potentially uncover unexpected relationships.A key element of EDA is visualisation, as through graphical presentations, users can intuitively understand trends, clusters, and correlations in the data. Communicating results through visualisation is often more succinct and efficient than just numbers[2].

In this thesis, Principal Component Analysis (PCA) has been chosen as the main tool for EDA. PCA is an algorithm that projects high dimensional data into a low dimensional space, preserving the direction of the largest variance in the data and reducing the dimensionality, thus visualising the data and making it easy for us to observe potential correlations in the data. Through the PCA score plot, we can visualise the similarity or difference between different fermentation samples, thus providing important hints for following analyses.

2.2.2 Supervised learning

Supervised learning is another large class of methods in machine learning, mainly used to learn predictive models from labelled datasets and to classify or regress new data. Common supervised learning algorithms include linear regression, support vector machines (SVM), decision trees, neural networks, and partial least squares discriminant analysis (PLS-DA). Supervised learning is trained by mapping relationships between input data (independent variables) and output labels (dependent variables), with the goal of finding models that best predict the labels of new data.

In my thesis, I chose PLS-DA (Partial Least Squares Discriminant Analysis) as the supervised algorithm to fit my data.PLS-DA is an algorithm commonly used for high-dimensional, multivariate data, and is especially suited for working with data that are highly correlated between the variables, which is typical of HS-GC-IMS data.PLS-DA is able to effectively dimensionality reduction of the data. At the same time, the VIP (Variable Importance Projection) of PLS-DA can extract the most important variables for classification, which provides the basis for subsequent chemical interpretation[2]. Therefore, PLS-DA can help HS-GC-IMS data to identify the key components of chemical changes during fermentation.

2.3 Persistent Homology

Persistent Homology (PH) is a topological tool for analysing the structure of data and can help us understand the shape and important features in the data. Unlike traditional analysis methods, PH focuses on the topological changes in data at different scales, especially those of persistent structures, and is unique in its ability to track features such as connectivity in data, which are "born" or "die" as a result of a change in a parameter (e.g., filtering threshold)[17]. PH evaluates the persistence of these topological features by recording their "birth" and "death". Persistent features mean that they are significant at multiple scales and therefore usually represent meaningful signals, while features that appear briefly may be noise or irrelevant information.

The data generated by HS-GC-IMS often have a complex two-dimensional structure containing two dimensions, retention time and drift time, as well as accompanying noise and background signals. Conventional peak detection methods may have difficulty in effectively distinguishing the true signal from the noise, especially in the case of high background signal interference. PH, on the other hand, is able to automatically extract the most persistent features from the topology of the data, helping to identify the true

signal peaks in a complex noise environment.

Hadi Parastar [17] successfully applied the PH algorithm to a single sample in his study. But HS-GC-IMS generates a large number of datasets with different peaks for each data. It becomes a challenge to apply this technique to huge datasets and to select meaningful peaks in the dataset. To solve this problem, this thesis combines this technique with the VIP score of PLS-DA, which will efficiently and accurately select the significant peaks in the dataset.

Chapter 3

Research methodology

This chapter presents the research methodology and experimental steps used in this thesis. From the collection and pre-processing of data, to the use of machine learning techniques to evaluate and analyse the data. By describing these processes in detail it provides the basis for the following analysis of the results. The data used in this thesis were provided by **Professor Bordiga from the Universiá del Piemonte Orientale**.

3.1 Preparation of Culture Medium and Sampling

3.1.1 Preparation of Culture Medium

The culture medium utilized was the classical sterilized MRS broth (Condalab, Madrid, Spain). Fermentation was conducted in three different set-ups simulating possible contamination. The bacterial strain used was *Lactobacillus paracasei* ATCC 6134, at an initial concentration of 1×10^6 CFU. Contamination was performed using a classical *Saccharomyces cerevisiae* (for bakery use) at a final concentration of 100 mg/L and a third contamination set-up where fermentation was conducted without sterility (no microbiological fume hood or sterile equipment).

- **MRS**: *L. paracasei* (1×10^6 CFU)
- **MRS + SC**: *L. paracasei* (1×10^6 CFU) + *S. cerevisiae* (100 mg/L)
- **MRS + X**: *L. paracasei* (1×10^6 CFU) + no sterility

Fermentation was carried out in a final volume of 100 mL in Erlenmeyer flasks. The flasks were placed in a shaking incubator at 37 °C, 200 rpm.

3.1.2 Sampling

During the experiments, sampling was carried out under a fume hood at 0 h, 2 h, 4 h, 6 h, and 24 h. For each sampling, 1 ml of the sample liquid was placed in a 20 ml glass vial using aseptic handling for subsequent HS-GC-IMS analysis. As shown in the table3.1, four different media and their combinations were used in the experiments, namely, a

single medium (MRS), a medium with Lactobacillus Paracasei (MRS + L), a medium with Lactobacillus Paracasei and Saccharomyces cerevisiae (MRS + L + S), and medium with the addition of Lactobacillus Paracasei and contaminated swabs (MRS + L + X). All samples were taken at the same time point to ensure comparability and accuracy of the experimental data.

Sample	Sampling	Code
Culture Medium	0, 2, 4, 6, 24 h	**(MRS)**
Culture Medium + *Lactobacillus Paracasei*	0, 2, 4, 6, 24 h	**(MRS + L)**
Culture Medium + *Lactobacillus Paracasei* + *Saccharomyces cerevisiae*	0, 2, 4, 6, 24 h	**(MRS + L + S)**
Culture Medium + *Lactobacillus Paracasei* + Contaminated swab	0, 2, 4, 6, 24 h	**(MRS + L + X)**

Table 3.1: Sampling Time Points and Codes for Various Culture Medium Conditions

3.2 HS-GC-IMS analysis

The HS-GC-IMS analysis of the fermentation samples was conducted using a GC-IMS instrument (FlavourSpec®, Dortmund, Germany). In order to ensure the accuracy of the experimental results, the parameters of the instrument were finely set and adjusted. Figures 3.1 and 3.2 show the workflow of the instrument and the specific parameters set, respectively.

As shown in Figure 3.1, the samples entered the gas chromatography column (GC Column) through the injection port, after which they were separated and analysed by ion mobility spectrometry (IMS). To ensure the accuracy of the analysis, nitrogen (N_2) was used as the carrier gas and drift gas, and the relevant temperature and pressure parameters were adjusted by electronic pressure controllers (EPC1 and EPC2). The heaters (T1-T5) labelled in the figure were used to heat the different experimental components to ensure proper temperature conditions.

Figure 3.2 shows the specific parameter settings used during the experiment. The drift gas flow rate (EPC1) was set to 75.0 ml/min, the carrier gas flow rate (EPC2) was set to 2.0 ml/min, and for the temperature settings, the temperatures of the IMS, the GC, the injection port, and the transfer tube were set to 45 °C, 45 °C, 80 °C, and 45 °C, respectively.These parameters were optimised in the pre-experiments to ensure the stability of the experimental process and the reliability of the data.

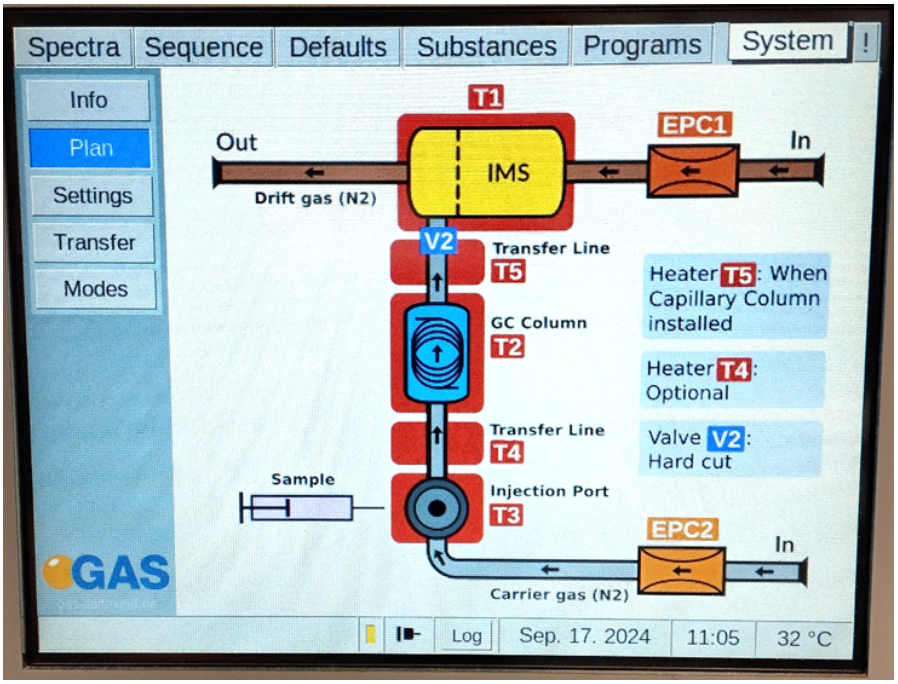


Figure 3.1: The workflow of the GC-IMS

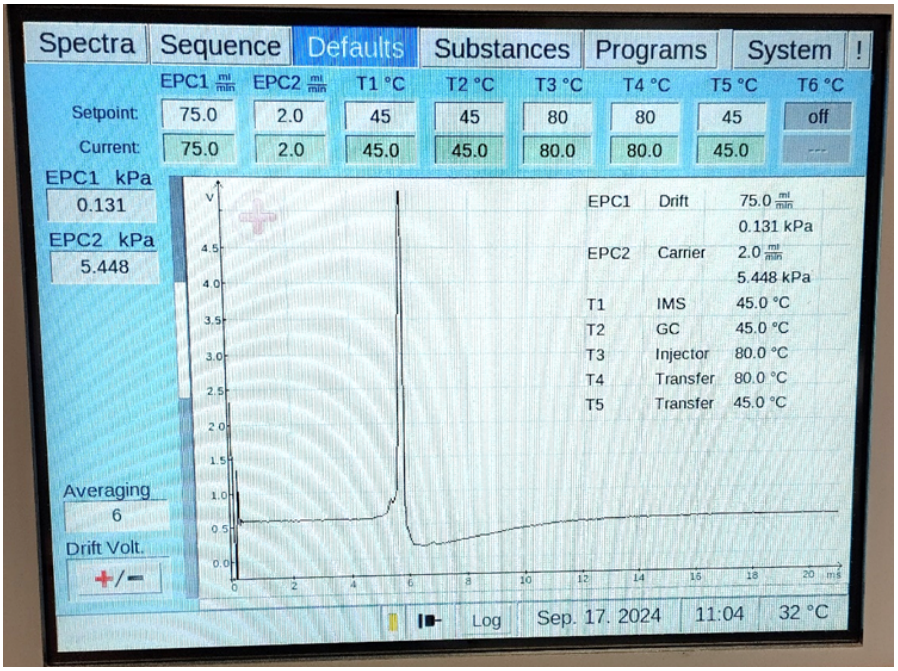


Figure 3.2: The specific parameters of GC-IMS

3.2.1 Headspace Sampling Conditions

- **Sample Preparation**: 1 mL fermentation samples were put into a 20 mL headspace glass sampling vial.
- **Incubation**: Incubate the sample at 40 °C for 5 minutes with a shaking speed of 500 rpm.
- **Injection**: The injection needle temperature is 80 °C, and the injection volume is 300 µL.

3.2.2 GC Conditions

- **Column**: MXT-5; low-polar capillary column (15 m × 0.53 mm, 1 µm film thickness).
- **Column Temperature**: 45 °C.
- **Analysis Time**: 10 minutes.
- **Carrier Gas**: Nitrogen (purity ≥ 99.999%).
- **Flow Rate**:
  - 0–3 min: 2 mL/min.
  - 3–7 min: 2–150 mL/min.
  - 7–10 min: 150 mL/min.

3.2.3 IMS Conditions

- **Drift Tube Length**: 5 cm.
- **Drift Tube Voltage**: 400 V/cm.
- **Drift Gas**: Nitrogen (purity ≥ 99.999%) at a flow rate of 75 mL/min.
- **IMS Temperature**: 45 °C.

3.3 Software

GC-IMS data export was performed using Vocal software (G.A.S.,Germany). After exporting the data, we used Python (version 3.11.2) to handle the subsequent processing and analysis, utilizing various libraries such as `gc-ims-tools` (version 0.1.7)for key pre-processing tasks. It is a package specifically designed for GC-IMS data. There are two important classes in it. The first one is the Spectrum class, which is used to represent and process individual spectra, containing all methods for input/output, preprocessing (such as cutting axes and visualization). The other is the Dataset class, which is used to manage multiple GC-IMS spectra, including sample names and labels. Methods that require multiple spectra, such as alignment or calculation of averages, are implemented here. Additionally, it provides tools for indexing, selecting, or deleting

samples, ensuring compatibility with other data science libraries, mainly using NumPy array formats. This package was also crucial for exploratory data analysis (PCA) and Supervised learning(PLS-DA), allowing for the rapid generation of heatmaps, PCA score plots, and VIP scores of the samples, clearly illustrating the differences between them. This laid the foundation for subsequent machine learning analyses[2].

For the post-preprocessing results, the following libraries were utilized:

`NumPy`: Used for efficient array computations, particularly effective in handling large GC-IMS datasets.

`Pandas`: Employed for data input/output (I/O) operations and manipulation of tabular data.

`Matplotlib and Seaborn`: Used for data visualization, aiding in the clear representation of the analysis results.

3.4 Data Pre-processing

We exported all HS-GC-IMS data as .mea files. In these files, each sample contains three main data matrices: a two-dimensional intensity matrix, a one-dimensional matrix for retention times (RTs), and another one-dimensional matrix for drift times (DTs). The size of the RTs matches the number of rows in the intensity matrix, while the size of the DTs corresponds to the number of columns in the intensity matrix, as shown in Table3.2. These data can also be used to generate a plot as figure 3.3 for further analysis.

	Drift time[ms]	0	6.66E-03	1.33E-02		20.98	20.986	20.993
Retention time [s]	**Intensity**	0	1	2	...	3147	3148	3149
0	0	1737	1505	1289	...	379	381	383
0.147	1	1723	1493	1278	...	375	375	374
2.94	2	1733	1501	1283	...	373	374	374
	...	...	...	...	...	...	...	...
599.613	4079	1726	1495	1278	...	364	364	363
599.76	4080	1721	1489	1271	...	368	370	372
599.907	4081	1718	1486	1270	...	368	369	370

Table 3.2: Data structure of MRS 0h.mea file

Analysing this data involves a number of steps, which I have divided into two main parts pre-processing and peak selection. As shown in the figure1.1, pre-processing usually includes data compression, RIP processing and alignment, selection of regions of interest, denoising, baseline correction, scaling and normalisation. After preprocessing we use PCA score plots to analyse the performance of our pre-processing results. Finally we put the data into PLS-DA model to train it and select the important peaks by getting VIP scores.

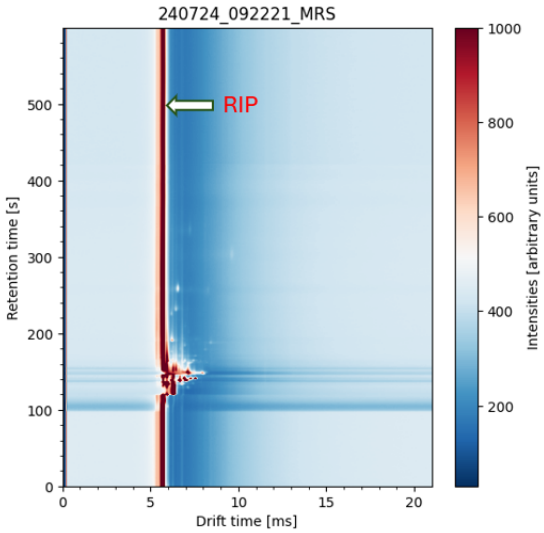


Figure 3.3: GC-IMS chromatogram of raw data (MRS 0h)

3.4.1 Binning and Drift time alignment

Binning

Binning techniques are widely used in food chemistry and play an important role in data pre-processing, especially when complex spectral data are involved, binning can greatly simplify data[14].

After we extract the intensity matrix from .mea, binning reduces the size of the matrix by combining every 2*2 neighbouring pixel points in the matrix into one value (their average). In our data from table3.2 we can see that the original data size is 4082*3150, after we applied binning the intensity matrix is reduced to 4082/2*3150/2(2041*1571). Thus the number of features is reduced to a quarter of the original.

Drift time alignment

After binning the data, I used the alignment of the drift times with the RIP as the reference signal. In subsection 2.1.3, when introducing the working principle of IMS, we introduced the concept of RIP that is a stable peak formed by the accumulation of hydrated hydrogen ions in the drift region.

There are several reasons why drift times need to be aligned. From the viewpoint of chemical analysis, after the alignment of each sample, the same chemical compounds will exhibit consistent drift times in different samples, avoiding the peaks from being misaligned during subsequent analyses. From the data analysis point of view, drift time alignment reduces the dimension of the data and thus improves the computational efficiency; on the other hand, since the RIP is represented in the intensity matrix as a high intensity value within a certain drift time interval, as shown in figura 3.3, these high values usually correspond to hydrated hydrogen ions and not to the compounds

that we are interested in. Therefore, this part of the data needs to be removed. After drift time alignment, the high intensity values of RIP may be reduced to a few columns or even disappear completely, thus reducing unnecessary interference and improving the accuracy and reliability of the subsequent algorithms. In summary, using RIP as a reference signal to align the drift time not only improves the consistency of the data, but also significantly enhances the accuracy and efficiency of data analysis.

To align the drift time based on the RIP, I performed the following steps:

1. Searched for the index of the maximum intensity along the drift time axis in the intensity matrix for each sample.
2. Took the median of all peak indices as the position of the RIP index.
3. Calculated the average drift time at this index position to determine the drift time of the RIP.
4. Used the formula (3.1) to normalize the drift time, obtaining a drift time relative to the RIP. This ensures that the drift time of each sample is aligned relative to the RIP.
$$t_{\text{rel}} = \frac{t_{\text{drift}}}{t_{\text{RIP}}} \tag{3.1}$$
5. Create a new drift time axis using cubic spline interpolation.

   After normalization, the drift times of different samples are relatively consistent, but the data points of each sample may still have slight differences, meaning that the sampling points of the drift time are not completely identical. To solve this problem, I established a new drift time axis and aligned all samples to this new axis. In order to map the intensity values to the new drift time axis, I used cubic spline interpolation to recalculate the intensity values for each sample at the same time points. This method provides a smooth and continuous way to interpolate these new time points, ensuring that the data is not distorted during the alignment process.

   In cubic spline interpolation, the relationship between intensity $I(t_{\text{rel}})$ and drift time t_{rel} is represented through the interpolation function (3.2).
$$I(t_{\text{rel}}) = S(t_{\text{rel}}) \tag{3.2}$$
   where $S(t)$ is the interpolation function, constructed based on the given relative drift time t_{rel} and the corresponding intensity values $I(t_{\text{rel}})$.

   The function $S(t)$ is a piecewise polynomial function. Each segment can be written as:
$$S_i(t) = a_i(t - t_i)^3 + b_i(t - t_i)^2 + c_i(t - t_i) + d_i \quad \text{for } t_i \leq t \leq t_{i+1} \tag{3.3}$$
   To project the intensity values $I(t_{\text{rel}})$ of each sample onto a unified drift time axis t_{new}, we need to perform interpolation calculations.

The formula is:

$$I(t_{\text{new}}) = S(t_{\text{new}}) = \sum_i \Big[a_i(t_{\text{new}} - t_i)^3 + b_i(t_{\text{new}} - t_i)^2 + c_i(t_{\text{new}} - t_i) + d_i\Big] \quad (3.4)$$

where a_i, b_i, c_i, d_i are determined through the following conditions[8]:
1. Interpolation Conditions
For each cubic spline $S_i(t)$ in the interval $[t_i, t_{i+1}]$, the spline must pass through the given data points (t_i, I_i) and (t_{i+1}, I_{i+1}). This leads to two interpolation conditions:

$$S_i(t_i) = I_i \quad \text{and} \quad S_i(t_{i+1}) = I_{i+1}$$

These conditions provide two equations for the spline coefficients.
2. First Derivative Continuity
To ensure smooth transitions between adjacent spline segments, the first derivatives of the splines must be continuous at the internal points t_i:

$$S_i'(t_i) = S_{i+1}'(t_i)$$

3. Second Derivative Continuity
For further smoothness, the second derivatives of the adjacent splines must also be continuous at each internal point t_i:

$$S_i''(t_i) = S_{i+1}''(t_i)$$

4. Boundary Conditions
Two common boundary conditions are used to finalize the system of equations:

- Natural Boundary Condition: The second derivative at the endpoints t_0 and t_n is set to zero, i.e.,

$$S_0''(t_0) = 0 \quad \text{and} \quad S_n''(t_n) = 0$$

- Clamped Boundary Condition: The first derivative at the endpoints is specified, e.g.,

$$S_0'(t_0) = f'(t_0) \quad \text{and} \quad S_n'(t_n) = f'(t_n)$$

After the above processing steps, I aligned all the feature values along an axis named "Drift time RIP relative." Since the RIP peak is located at the beginning of this new axis, I selected the range between 1.05 and 2.5 to remove the RIP signal. This range approximately corresponds to a drift time of 6 to 21 milliseconds.

3.4.2 Calculate Means and Region of interest (ROI) selection

Calculate means

By averaging the intensity matrices of all the samples, I was able to obtain a representative intensity matrix that would contain common features across all the samples. Averaging reduces noise interference and thus highlights those feature peaks that occur consistently in the region of interest. In this way, it is guaranteed that the region of interest is selected to cover the characteristic points of all samples.

Let I_i represent the intensity matrix of the i-th sample, and let N be the total number of samples. The average intensity matrix $\overline{I}$ across all samples can be expressed as:

$$\overline{I}(x, y) = \frac{1}{N} \sum_{i=1}^{N} I_i(x, y)$$

I have presented the obtained mean intensity matrix in the form of a Spectrum as in Figure 3.4. By comparing Figure3.4 with the plot of sample MRS 0h (Figure 4.2), we can see that the features contained in Figure 3.4 are significantly larger than those in Figure 4.2. That is to say, we have all the features of the samples represented in the intensity matrix of the mean values.

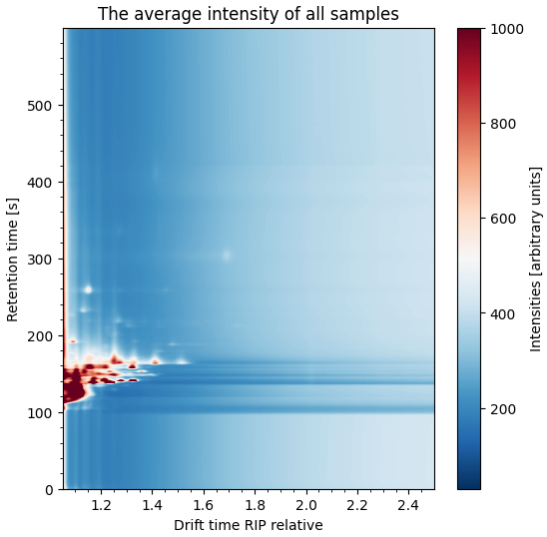


Figure 3.4: GC-IMS chromatogram of the average intensity matrix

Region of interest (ROI) selection

In order to achieve automation the ROI selection process. First I use the iterative adaptive thresholding algorithm to compute a threshold that distinguishes between the foreground and background of the chromatograms of the mean intensity matrix.Then a binarized matrix is obtained by binarising the intensity matrix based on the calculated threshold. Finally the ROI is selected based on the binarised matrix.

The steps of the iterative adaptive thresholding algorithm are as follows:

- Initialize the threshold T_0 as the average of the maximum and minimum pixel values in the image:

$$T_0 = \frac{\max(I(x,y)) + \min(I(x,y))}{2} \tag{3.5}$$

  where $\max(I(x,y))$ and $\min(I(x,y))$ represent the maximum and minimum intensity values of the image $I(x,y)$.

- Define two regions, A and B, based on the current threshold T_n:

$$A = \{I(x,y)) \mid I(x,y)) \geq T_n\} \tag{3.6}$$

$$B = \{I(x,y)) \mid I(x,y)) < T_n\} \tag{3.7}$$

- Compute the mean intensity values for regions A and B, denoted as g_A and g_B, respectively:

$$g_A = \frac{1}{|A|} \sum_{(x,y)\in A} I(x,y)) \tag{3.8}$$

$$g_B = \frac{1}{|B|} \sum_{(x,y)\in B} I(x,y)) \tag{3.9}$$

  where $|A|$ and $|B|$ are the number of pixels in regions A and B.

- Update the threshold to the average of g_A and g_B:

$$T_{n+1} = \frac{g_A + g_B}{2} \tag{3.10}$$

- The iterative process continues until the difference between the current threshold T_n and the previous threshold T_{n+1} is smaller than a small constant ϵ:

$$|T_{n+1} - T_n| < \epsilon \tag{3.11}$$

  where ϵ is a predefined tolerance value that controls the convergence precision. I chose it to have a value equal to 0.00001, in order to ensure the accuracy of the threshold.

- Once the threshold has converged, the final threshold T is returned as:

$$Th = T_{n+1} \tag{3.12}$$

In the second step of automating the ROI selection process, I used the final threshold to construct a binarized matrix, where the regions containing 1 represent the ROI. The task was to extract all the regions in the matrix that contain 1. First, by scanning each column, we identified the last column that contains a 1 and recorded its index as index_dts_end. Since the drift time alignment has been completed, there is no need to process the left part of the matrix further. Next, by scanning each row, we found the first and last rows containing 1, and recorded their indices as index_rts_start and index_rts_end, respectively. Finally, based on these indexes, we sliced the intensity matrix to obtain our ROI.

3.4.3 Denoising and Baseline correction

Denoising

In addition to being used for ROI selection, the threshold calculated in the previous step serves another crucial purpose: denoising, particularly in data with high noise levels, where noise can interfere with subsequent peak extraction. In this case, a threshold of 750 was automatically calculated and applied to all samples. Any intensity values below 750.8 were considered noise and set to 0 in the intensity matrix. This approach significantly reduces data volume, retaining only the values relevant to the actual signals, such as chemical feature peaks.

Baseline correction

The baseline is the signal level when no compounds are detected in the HS-GC-IMS, and it is usually a more stable signal. The baseline primarily reflects the background level of noise and represents the minimum response of the detection system. When there is no substance to be measured, the system still picks up some ambient noise or background signals that make up the baseline[16]. This baseline level may vary slowly due to fluctuations in temperature, pressure or the instrument itself in the experimental environment, so baseline correction is essential. Especially during long experimental periods, signal drift and instrument fluctuations can affect the stability of the baseline, resulting in a gradual increase in the difference between the detected signal and the baseline.

The main purpose of Baseline correction is to ensure that the peak signal ultimately obtained is the true signal of the target compound and not a false signal caused by noise or baseline drift. Uncorrected baselines may cause fake peaks to appear, which in turn can interfere with the analysis results. With baseline correction, the interference of ambient noise can be effectively eliminated and the true signal of the compound can be made clearer and more reliable, thus improving the accuracy of the data. A corrected baseline brings the peak signal closer to the actual concentration change of the compound, reducing the effect of noise on the data, and effective baseline correction ensures that each peak feature is accurately identified.

Asymmetric Least Squares (ALS) is a baseline correction technique primarily used for signals like spectra[16]. In GC-IMS data, there are often irregular or high-intensity peaks that require effective handling while ignoring noise below the baseline. The goal of ALS is to estimate a baseline that is smoother than the observed signal y, while preserving the essential peak features.

The ALS optimization function is defined as:

$$s = \sum_{ij} \omega_{ij} d_{ij}^2 + \lambda \sum_{ij} (\Delta^2 z_{ij})^2 \tag{3.13}$$

- The first term measures the deviation between the observed signal y_{ij} and the estimated baseline z_{ij}

$$\sum_{ij} \omega_{ij} d_{ij}^2$$

where $d_{ij} = y_{ij} - z_{ij}$

- The weight ω_{ij} controls the influence of the peak signal on the baseline correction and is defined as:

$$\omega_{ij} = \begin{cases} p & \text{if } d_{ij} > 0 \\ 1 - p & \text{otherwise} \end{cases}$$

- where p is typically set between 0.001 and 0.1. By adjusting the weight p, the influence of the peaks on the baseline fitting can be controlled, preventing overcorrection that might distort the signal.
- The second term is a smoothness penalty to ensure the baseline z_{ij} is continuous and smooth

$$\lambda \sum_{ij} (\Delta^2 z_{ij})^2$$

.

- The smoothness term is given by:

$$\Delta^2 z_{ij} = z_{ij} - 2z_{ij-1} + z_{ij-2}$$

This term ensures smooth transitions in the baseline between consecutive points.

- The parameter λ, which controls the degree of smoothness, ranges from 10^2 to 10^9. A larger λ results in a smoother baseline. By adjusting λ, ALS can balance between fitting the baseline tightly to the observed data and maintaining peak integrity.

The key to ALS correction is balancing the deviation between the signal and baseline while ensuring smoothness in the baseline. This method is particularly effective in handling spectral data with irregular peaks or high-intensity noise, allowing for accurate extraction of chemical signal features by reducing baseline drift.

3.4.4 Scaling

Before applying PCA (Principal Component Analysis) and PLS-DA (Partial Least Squares Discriminant Analysis), data scaling is necessary because different variables may have different units and ranges. Without scaling, variables with larger numerical values may dominate the results, while those with smaller values may be overlooked. By scaling the data, we ensure that each variable is analyzed on the same scale, allowing PCA and PLS-DA to more accurately capture the underlying structure and characteristics of the data.

The scaling formula is as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{3.14}$$

Where:
-x is the original data value
-μ is the mean of the variable

-σ is the standard deviation of the variable
-x' is the scaled data value
Through this formula, a variable with different scales can be transformed into a standard normal distribution with a mean of 0 and a standard deviation of 1.as shown in figure3.5.. In this way, the effects of variables with different scales on the results of PCA and PLS-DA can be avoided, and the analysis can be ensured to be fair and accurate.

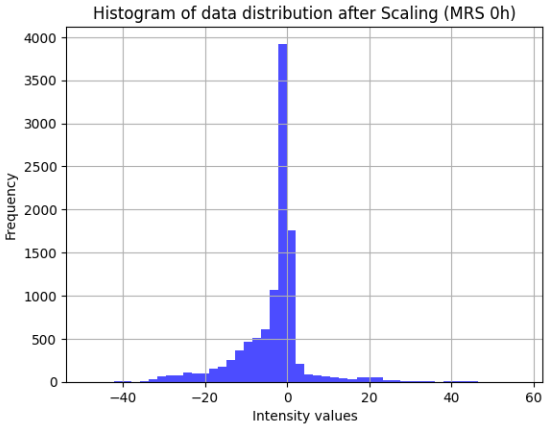


Figure 3.5: Histogram of data distribution after Scaling (MRS 0h)

To make the data available for analysis by machine learning algorithms, after scaling the data, I need to construct a matrix X to ensure that all the samples have aligned features and have the same dimensions.
The process involves the following key steps:

- **First**, convert each sample's two-dimensional intensity matrix into a one-dimensional vector. This means flattening the sample's features (e.g., ion drift time and retention time intensity values) into a single row, with each row representing all the features of a sample. This process retains the original feature information, preparing it for further analysis.

- **Next**, stack the feature rows of all samples together to form a complete X matrix. Each row of matrix X represents a sample, and each column represents a specific feature. This ensures that the features of all samples are aligned consistently and that each sample has the same number of features, avoiding dimensional inconsistency issues.

- **Finally**, the resulting matrix X has dimensions $N \times J$, where N is the number of samples and J is the number of features per sample. The X matrix is then used as input for machine learning algorithms, such as PCA (Principal Component Analysis) or PLS-DA (Partial Least Squares Discriminant Analysis).

3.4.5 Evaluation of pre-processing results in combination with PCA

As you can see from Table 3.1 that each of our data is grouped and labeled, I wanted to use visualization tools in EDA to evaluate whether or not the results of the pre-processing are somehow related to the data labels. The reason for choosing the score plot of the PCA algorithm to evaluate the pre-processing results is that PCA is able to extract and project the main variation information in the data into a low-dimensional space, usually two or three dimensions, through dimensionality reduction. Through this projection, score plots are able to visualize similarities and differences between samples. Ideally, the pre-processing result should make the relationship between samples more explicit, e.g., similar samples should be clustered together, while more differentiated samples should be spread far apart. Therefore, score plots can be used as a visual tool to assess the effectiveness of pre-processing and to verify that the preprocessing steps have improved the quality of the data by observing whether the data points form a reasonable clustering or distribution pattern.
The steps to get the PCA score plot are as follows:

1. The input data matrix X is an $N \times J$ matrix, where:

   - N is the number of samples,
   - J is the number of features.

   To capture the relationships between different features, we calculate the covariance matrix C, which describes the linear relationship between each pair of features. The covariance matrix is calculated as follows:

   $$C_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \mu_j)(X_{ik} - \mu_k) \tag{3.15}$$

   where:

   - X_{ij} and X_{ik} are the j-th and k-th feature values of the i-th sample,
   - μ_j and μ_k are the means of the j-th and k-th features,
   - C_{jk} represents the covariance between the j-th and k-th features.

   By calculating the covariance for all feature pairs, we obtain a $J \times J$ covariance matrix C. The simplified computation for the covariance matrix is:

   $$C = \frac{1}{N-1} X^T X \tag{3.16}$$

   where X^T is the transpose of the input data matrix X.

2. The covariance matrix is decomposed to obtain eigenvalues λ and eigenvectors v. Eigenvalues and eigenvectors provide important information about the variance in the data. Eigenvalues represent the amount of variance explained by each

principal component, while eigenvectors represent the direction of the principal components[15].

The eigenvalue decomposition is expressed as:

$$\det(C - \lambda I) = 0 \tag{3.17}$$

$$(C - \lambda I)v_i = 0 \tag{3.18}$$

where:

- λ is the eigenvalue of the covariance matrix C,
- I is the identity matrix,
- v_i is the eigenvector corresponding to the eigenvalue λ_i.

Larger eigenvalues indicate that the corresponding principal component explains more variance in the data.

3. Typically, we select the eigenvectors corresponding to the largest eigenvalues as the principal components (PCs). By selecting a few principal components, we can effectively reduce the dimensionality of the data while retaining most of the important information[15].

4. After determining the principal components, we project the original data onto these components to calculate the scores for each sample. Scores represent the position of each sample in the principal component space, and the calculation is as follows:

$$Score_{x,i} = X \cdot v_i \tag{3.19}$$

where:

- X is the original data matrix,
- v_i is the eigenvector corresponding to the i-th principal component.

This formula projects the data matrix X onto the principal components, generating a score matrix S, where each row represents the coordinates of a sample in the principal component space.

5. Finally, the score plot helps us visualize the distribution of samples in the principal component space. By observing the score plot, we can identify clustering or separation between samples, revealing patterns or structures in the data.

3.5 Peak selection

This section discusses the peak selection methods employed in HS-GC-IMS data analysis. Peak selection is a critical step in extracting important features from GC-IMS chromatogram data that can help provide deeper knowledge of the chemical composition of the sample. Several methods can be used to identify and select relevant peaks, including persistent homology, uncertainty analysis, and multivariate analysis techniques (e.g., PLS-DA).

3.5.1 Topological data analysis

Topological Data Analysis (TDA) is a method for extracting information about the whole structure and shape from complex, high-dimensional data. By analyzing the geometric features of data, TDA helps us discover patterns and trends that are difficult to identify with traditional statistical methods[17]. In practical applications, data often contains some kind of geometric shape that carries important information in the data. For example, in a topographic map, topographic features such as mountains and valleys can be extracted by analyzing the geometric structure of the data.

3.5.2 Persistent Homology

One of TDA's core tools is Persistent Homology(PH), which is designed to quantify topological features in different dimensions of the data and determine their significance by analyzing the 'persistence' of these features[17]. In PH, a *simplicial complex* is a fundamental tool used to represent the geometric and topological structure of data. It provides a way to discrete complex geometric spaces, enabling the analysis of topological features within data using computational algorithms[1].

A *simplex* is the basic building block in geometry. An n-dimensional simplex is the convex hull of $n+1$ points, with the most common examples being[11]:

- 0-simplex: a point, representing an individual point.
- 1-simplex: a line segment, formed by two points.
- 2-simplex: a triangle, formed by three points.
- 3-simplex: a tetrahedron, formed by four points.

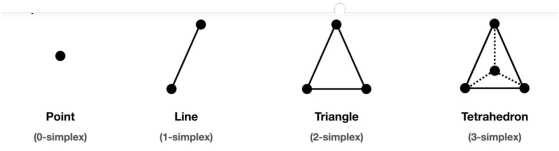


Figure 3.6: N-dimensional simplex[20]

The dimension of a simplex is determined by the number of vertices it has, and each n-dimensional simplex's face is an $(n-1)$-dimensional simplex, as shown in figure3.6.

For instance, the faces of a triangle are line segments, and the faces of a tetrahedron are triangles.

A *simplicial complex* is a collection of simplices that satisfies the following two properties[1][11]:

- If a simplex is in the complex, all its faces must also be included in the complex.
- If two simplices intersect, their intersection must be a common face of both simplices.

This structure allows us to build complex topological spaces using simple geometric objects, which can be used to analyze features like connectivity and holes within the data.

In PH, simplicial complexes are utilized to convert data points into computable topological structures, as shown in figure3.7. By connecting points, line segments, triangles, and other simplices into a simplicial complex[1], Algorithm can capture essential topological features of the data, such as:

- Connected components: distinct regions or clusters in the data (captured by 0-dimensional simplices).
- Holes: enclosed but unfilled areas in 2D space (captured by 1-dimensional simplices).
- Tunnels or cavities: voids or tunnels in 3D space (captured by 2-dimensional simplices and higher).

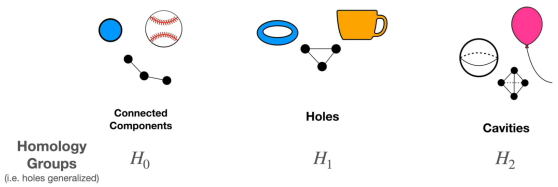


Figure 3.7: Simplicial complexes[20]

Persistent homology analyzes the topological features of simplicial complexes over a filtration process, tracking the birth and death of these features and quantifying their persistence, which helps identify the most significant shape patterns within the data.

3.5.3 Persistence diagram

Persistence Diagram (PD) is an important tool for Persistent Homology (PH) technique in topological data analysis, which records the process of generation and disappearance of topological features as a function of scale, thus revealing the deep structural information of the data.

Each data point in the GC-IMS data can be regarded as a 0-simplex (point), which constitutes a simplex complex by connecting neighboring points to generate 1-simplexes (line segments).The simplex complex gradually expands and changes as the data gradually decreases the signal strength (Values of the intensity matrix) threshold. In this process,

different topological features (e.g., connectivity components) are 'born' or 'die'. Specifically, 'birth' refers to the scale at which the connected component first appears, while 'death' is the scale at which the feature merges with other features or disappears[17]. Let us take the sea as an example, as shown in figure3.8. The water level at an altitude of 300 is constantly decreasing. At the highest localised point, islands(0-simplex) appear (birth). As the water level decreases (intensity matrix threshold decreases), two islands are connected (connected component): the lower island is connected to the higher island (death). The so-called persistence graph represents the death value of all islands compared to the birth value.

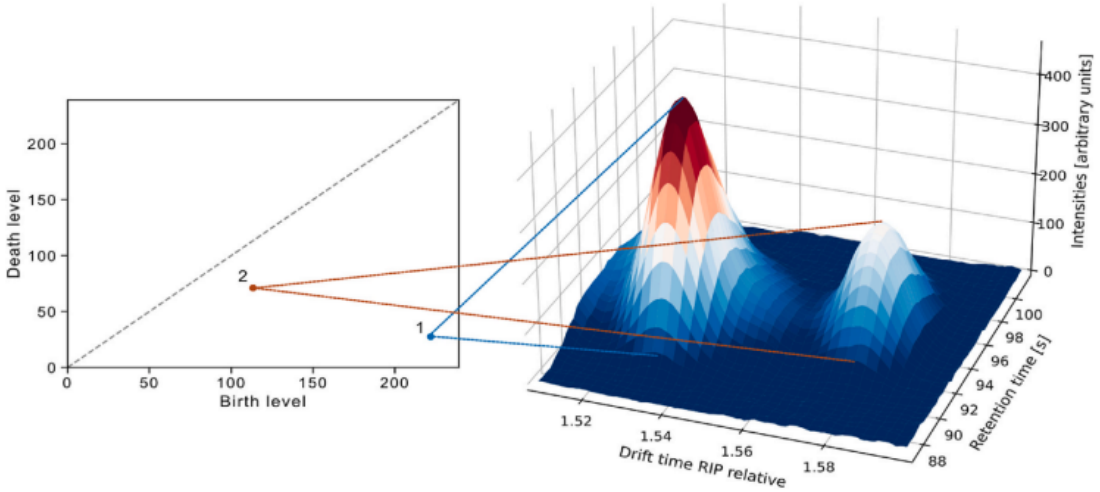


Figure 3.8: Persistence plot corresponding to the upper-level set filtration of GC-IMS data as an example.[17]

By calculating the birth and death scales of features, PH can determine the duration of each topological feature. Features with long durations are often considered important structures in the data, while features with short durations may be noise. The duration graph plots the duration of these topological features on a two-dimensional plane, where the x-axis represents the birth scale of the feature and the y-axis represents the death scale of the feature. Each feature is represented as a point in the figure3.8, and the vertical distance of the point reflects its duration. Points that are farther away from the diagonal line indicate longer durations of features and are usually considered more stable or important. Continuity diagrams not only help analyze the robust topology in the data, but also enhance the understanding of the data by identifying persistent features.

3.5.4 Application of the PH algorithm

Before applying the persistent homology (PH) algorithm to generate the persistence diagram, several pre-processing steps were necessary to prepare the GC-IMS chromatogram data for analysis. These pre-processing steps included color scaling, conversion to

grayscale, and subsequent feature extraction using the PH algorithm. The following steps, described in detail, lay the foundation for building persistent charts.

- **Color Scaling and Grayscale Conversion:**

  The original GC-IMS chromatogram contains rich color information, with pixel intensities representing the strength of the detected signals. To facilitate further analysis, we first rescaled the color pixel values to ensure they fall within the range of [0, 255]. The following algorithm was applied to achieve this scaling:

$$Ximg' = Ximg \times \left(\frac{255}{\max(Ximg)} \right) \tag{3.20}$$

  Here, $Ximg$ represents the original pixel values of the image(intensity matrix), and $\max(Ximg)$ is the maximum pixel value in the image. This scaling adjusts all pixel values to a 0-255 range, making the color contrast more pronounced. Enhancing the contrast was crucial for correctly identifying features like peaks and valleys, as it allowed the PH algorithm to detect significant topological features more efficiently.

  Once the pixel values were rescaled, the next step was to convert the image to grayscale. Grayscale conversion simplified the image by reducing it to intensity values only, with brighter pixels representing higher intensities and darker pixels indicating lower intensities. This transformation made the image more suitable for topological analysis by removing the complexity introduced by color.

- **Peak Extraction Using the PH Algorithm:**

  After the image was rescaled and converted to grayscale, we applied the PH algorithm to extract significant features, specifically the peaks in the GC-IMS chromatogram. The PH algorithm detects topological features by tracking their "birth" and "death" times, effectively identifying regions where notable features such as peaks appear and disappear.

  For each identified peak, several key parameters were recorded:

  - **Drift time:** Corresponding to the horizontal axis of the chromatogram, representing the ion mobility spectrometry (IMS) drift time.
  - **Retention time:** Representing the vertical axis of the chromatogram, indicating the retention time during gas chromatography (GC) separation.
  - **Pixel position (x, y):** The coordinates of the peak in the image matrix.
  - **Birth level:** The grayscale intensity value where the peak begins to emerge.
  - **Death level:** The grayscale intensity value where the peak disappears.

- **Construction of the Persistence Diagram:** Using the birth and death values from the table, we proceeded to construct the persistence plot. In the figure, each point represents a topological feature (in this case, a peak and valley).

3.5.5 Uncertainty analysis of HS-GC-IMS

After completing the peak detection, the uncertainty analysis of HS-GC-IMS is performed to estimate and quantify the sources of error in the experiment and to ensure the accuracy and reliability of the results. Uncertainty analysis helps us to identify potential problems in the detection process, such as instrumental errors, environmental effects, and instabilities in data processing. By evaluating parameters such as drift time, retention time and intensity of each peak, the precision and confidence interval of the assay can be clarified. This process is essential to optimise the experimental process and improve data repeatability, as well as to ensure that statistically significant peaks are screened to reliably reflect the chemical composition of the sample.

For this purpose, Professor Bordiga provided a data set. This dataset consists of ten samples, each of which is a coffee roasted and brewed from the same variety of coffee beans. This dataset was chosen because the aroma compounds in coffee show a high degree of stability after the beans have been roasted. Therefore, this data set is well suited for assessing the measurement uncertainty of the instrument. By repeating measurements on these samples, possible random errors of the instrument under different experimental conditions can be more clearly identified.

Describe the calculation process in the following:

- **Data Preprocessing and Analysis:**

  In this study, for each coffee sample, we applied the same pr-eprocessing techniques described in section 3.4 as well as the methods detailed in section 3.5. During the peak selection process, each sample generated a corresponding peak table that included the drift time, retention time, and the coordinates of the peaks in the matrix. The results were then visualized using GC-IMS chromatograms, as shown in the figures 3.9.

- **Comparison of Peaks Across Samples:**

  By comparing the peak positions across the ten samples in the chromatograms, it became clear that the locations of the first ten peaks were generally similar. To simplify the analysis, one sample was randomly chosen as a reference sample, and its peak positions were used as a reference value. The peak coordinates of the other nine samples were then compared against this reference.To ensure that peaks selected across different samples corresponded to the same physical peak, a threshold was defined: peaks within a distance of 10 pixels from the reference peak were considered to represent the same peak.

- **Standard Deviation Calculation:**

  After identifying corresponding peaks across samples, we calculated the uncertainty between these peaks using the formula for standard deviation. The standard deviation formula is as follows:

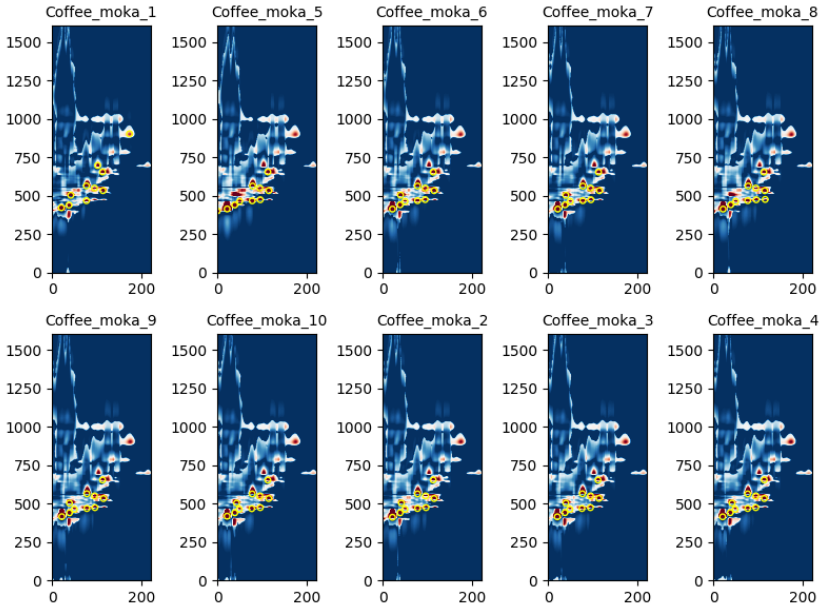


Figure 3.9: Top 10 significant peaks in all coffee samples

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{3.21}$$

where σ is the standard deviation, N is the number of samples, x_i is the value of the i-th sample, and μ is the mean of the sample values. Standard deviation measures the spread of the peak positions across samples; a higher standard deviation indicates greater variation and, thus, greater measurement uncertainty.

3.6 PLS-DA and VIP scores

Through the above techniques, we can successfully extract relatively significant peaks from a sample. However, to determine which peaks or features are important for the whole database, further analysis is required with the help of machine learning methods. In particular, we assess which peaks or features have a greater effect on the classification and prediction of a sample by training a machine learning model. In this way, it is possible to identify the features that are most discriminating to the samples. The ultimate goal of this task is to extract key peaks from the database and further analyse the pattern of change of these peaks during the fermentation process, in order to better understand and control the key factors in the fermentation process, and thus optimise the production process and product quality.

3.6.1 PLS-DA (Partial Least Squares Discriminant Analysis)

PLS-DA (Partial Least Squares Discriminant Analysis) is one of the commonly used algorithms in GC-IMS data analysis, which is particularly suitable for dealing with datasets with high dimensionality and a small number of samples. Since GC-IMS data usually have a much larger number of peaks than the number of samples, PLS-DA is able to extract the key features by dimensional reduction for effective sample classification. The advantage of this algorithm is its supervised learning property, which enhances the separation between different classes, allowing even small changes in chemical signals in the samples to be detected. Combined with appropriate data pre-processing methods, PLS-DA not only improves classification accuracy, but also enhances the interpretation of results, demonstrating excellence in applications such as fermentation process monitoring. The main idea of PLS-DA is to extract Latent Variables (LVs) that can differentiate different classes by finding the linear relationship between the input variable X and the output label Y[5]. These latent variables can help us to reduce the dimensionality of the data while maintaining as much separation as possible between the sample categories to achieve classification. We can obtain LVs from the following steps:

1. *Input Data*

   Input data matrix X $(N \times J)$ that obtained in section 2.4.4.
   - N : the number of samples
   - J : the number of features.
   Output matrix y $(N \times 1)$, represents the classification labels for each sample.

2. *Calculating Weight Vector w*
   $$w^T = X^T y \tag{3.22}$$
   - X^T : the transposed input matrix
   - y : the output vector.
   -The resulting weight vector w $(J \times 1)$, indicates the contribution of each feature to the model.

3. *Calculate and normalize scores vector t*
   $$t = \frac{Xw}{\sqrt{w^T X^T X w}} \tag{3.23}$$
   -w^T : the transposed wright vector
   - t $(N \times 1)$: A vector representing the projection of each sample in LV space

4. *Calculate the X loadings vectors*
   $$p = X^T t \tag{3.24}$$
   - p $(1 \times N)$: The coefficients to indicate the measure in which each input feature X contributes to each LV.

5. *Calculate the Y loading* Calculate Regression coefficient - q (1×1): The coefficients to indicate the misuse in which Y contributes to each LV.
   - y^T : the transposed Output matrix

6. *Update the X and y residuals vector*

$$\text{rex\_x} = X - tp \tag{3.25}$$

$$\text{res\_y} = y - tq \tag{3.26}$$

- The variables rex_x and rex_y replace the initial X and Y in the calculation of the following latent variable.
- Back to step 2 with rex_x and rex_y.

7. *Calculate Regression coefficient*

$$b = w\left(p^T w\right)^{-1} q^T \tag{3.27}$$

- b : The coefficients are used to construct the relationship between X and y.

To ensure model accuracy I chose n_LV= 3. Our goal is to find a set of LVs such that the covariance between X and Y is explained at most by a linear combination of these variables.

3.6.2 VIP (Variable Importance Projection)

The variable importance projection (VIP) scores generated by the PLS-DA model plays a pivotal role in the analyses.The VIP scores are used to assess the contribution of each variable to the model, allowing for the identification of chemical peaks or other feature variables that have the most discriminatory power in the classification task[4]. Specifically, a higher VIP score means that the corresponding feature has a greater impact on the model in the differentiation of classes, thus providing information on which features have a key role in the classification process of the samples. In addition, VIP scores not only summarise the contribution of each variable to all potential variables in the PLS-DA model, but also combine the regression coefficients and variables in describing response Y (i.e., the output category) variance in significance. Therefore, the VIP score becomes is a tool critical for feature selection and model interpretation. It effectively filters out redundant information and focuses on those features that have a significant impact on the classification results.

The VIP score for a given variable j is calculated using the following formula[4]:

$$VIP_j = \sqrt{J \cdot \frac{\sum_{a=1}^{A}\left(SSY_a \cdot \left(\frac{w_{ja}}{\|w_a\|}\right)^2\right)}{\sum_{a=1}^{A} SSY_a}} \tag{3.28}$$

Where:

- VIP_j is the VIP score for the j-th feature.
- J is the total number of features.
- A is the number of latent variables (LVs) in the model.

- SSY_a is the explained variance of the response variable y for the a-th latent variable.
- w_{ja} is the weight of the j-th variable for the a-th latent variable.
- $\|w_a\|$ is the norm of the weight vector for the a-th latent variable.

The explained variance for each latent variable a, denoted as SSY_a, can be computed using the following formula[4]:

$$SSY_a = b_a^2 \cdot (t_a' t_a) \tag{3.29}$$

Where:

- b_a is the regression coefficient for the a-th latent variable.
- t_a is the score vector for the X-variables corresponding to the the a-th latent variable.
- $t_a' t_a$ represents the sum of squares of the score vector, indicating the contribution of the component to the X-variables.

Each feature j has a corresponding weight w_{ja} for each latent variable a. This weight reflects how much the feature contributes to the latent variable. The squared normalized weight $\left(\frac{w_{ja}}{\|w_a\|}\right)^2$ ensures that features with higher contributions to the latent variables receive higher VIP scores. The sum over all latent variables A ensures that the VIP score incorporates the contributions of the feature across all components of the model. Normalization by the total explained variance $\sum_{a=1}^{A} SSY_a$ ensures that the VIP scores are comparable across different features, even when the latent variables explain different amounts of variance.

VIP scores are a critical tool in PLS-DA for evaluating the importance of features in the model. By combining the contribution of each feature across all latent variables and considering the explained variance of the response y, VIP scores help identify the most influential variables for classification. This is particularly useful in high-dimensional datasets, where VIP scores can aid in feature selection, model optimization, and result interpretation. Variables with higher VIP scores should be considered crucial for the model's performance, while those with lower scores can potentially be excluded.

3.6.3 K-fold cross-validation method

Since PLS-DA is a classification algorithm, it is necessary to divide the dataset into a training set and a test set when using it. However, my main purpose of using the algorithm is to obtain VIP scores, and after directly dividing the dataset, I can only obtain the VIP scores of the training set part, and cannot extract the features of the test set. In order to solve this problem, I used the K-fold cross-validation method, and specifically chose K=5 for the 5-fold cross-validation.In each validation round, one sample from each sample group (for instance, one from a specific time point) is selected as the test set, while the remaining samples from the other four groups are used as the

training set. The process is repeated 5 times, rotating the sample selection such that each sample group is used as the test set once.

The steps for K-fold cross-validation are as follows:

1. For each fold, select 1 sample (e.g., from a specific time point) from the dataset as the test set.
2. Use the remaining 4 sample groups as the training set.
3. Train the model and calculate the corresponding VIP scores for each fold.
4. Rotate the sample selection so that each sample group is used as the test set once, ensuring balanced validation.

After completing the training for each fold, a set of VIP scores is generated. I then sum the VIP scores obtained from each fold, resulting in a comprehensive VIP score based on all samples. This approach not only ensures that the features from both the training and test sets are utilized, but it also provides a more accurate assessment of each variable's importance across the entire dataset. By using 5-fold cross-validation, I am able to overcome the limitation of only extracting VIP scores from the training set. This method ensures that VIP scores are generated based on the entire dataset, while maintaining the balance between the sample groups during each validation fold.

Chapter 4

Results and discussion

In this chapter I will show in detail the experimental results of the analytical study.

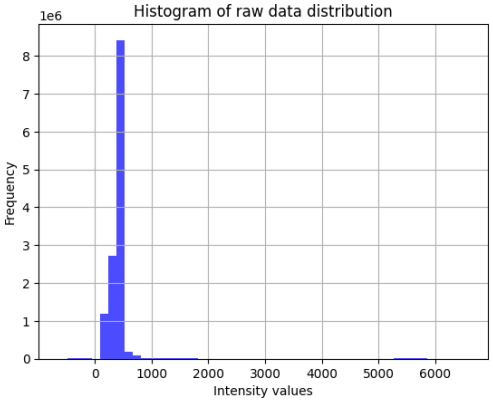
4.1 Results of Data Pre-processing

Firstly the results of each step in the pre-processing will be shown and discussed through images and tables. Then the results of the pre-processing will be analysed using the PCA score plot.

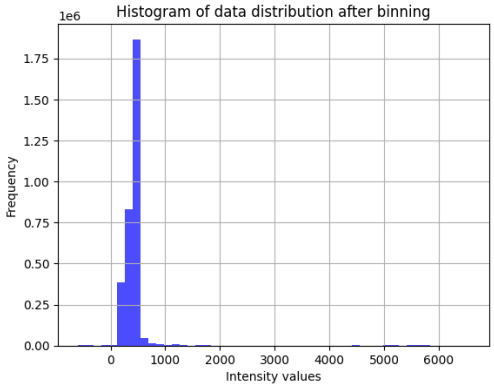
4.1.1 Results of binning and Drift time alignment

I plotted two histograms (figure 4.1) in order to better observe the changes in the distribution of the data after binning. The horizontal coordinates of the plots represent the values in the HS-GC-IMS intensity matrix, and the vertical coordinates represent the frequency of appearance of these values. In the left histogram4.1a, a large number of intensity values are concentrated between 0 and 1000. And the frequency (y-axis) reaches an extremely high magnitude, close to 8×10^6. This means that the intensity values of raw data has a highly concentrated in the lower range of intensities. In the right histogram (after binning) 4.1b, the whole shape of the data distribution keeps similar. It is still mainly concentrated between 0 and 1000, but the frequency values significantly decrease. The maximum frequency value is approximately 1.75×10^6. This shows that the binning approach effectively reduces the size of the intensity matrix while retaining the key features needed for further analysis.

After I aligned the drift times, the results are shown in the figure4.2.By comparing Figure 3.3, we can clearly see that the left part of the GC-IMS chromatogram of raw data was completely excised and the RIP was successfully removed. Then by observing the drift time axis, it can be seen that all the features in Figure 4.2 correspond to a new timeline relative to the RIP. This shows the successful application of the three-spline interpolation method to our data. It means that the same peaks in all samples correspond consistently at the time points.



(a) Histogram of raw data distribution



(b) Histogram of data distribution after binning

Figure 4.1: Comparison of raw data distribution and data distribution after binning

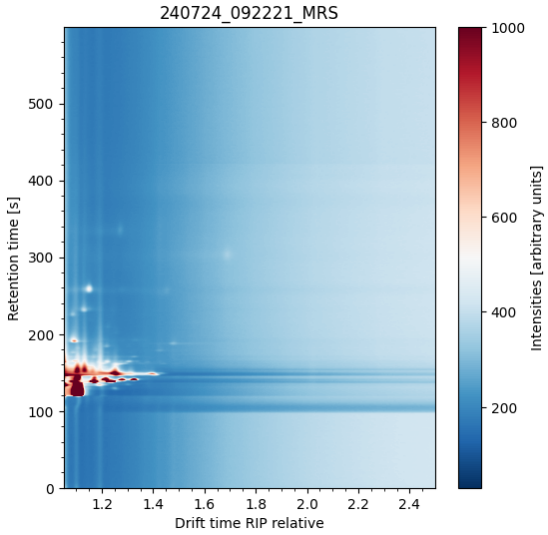
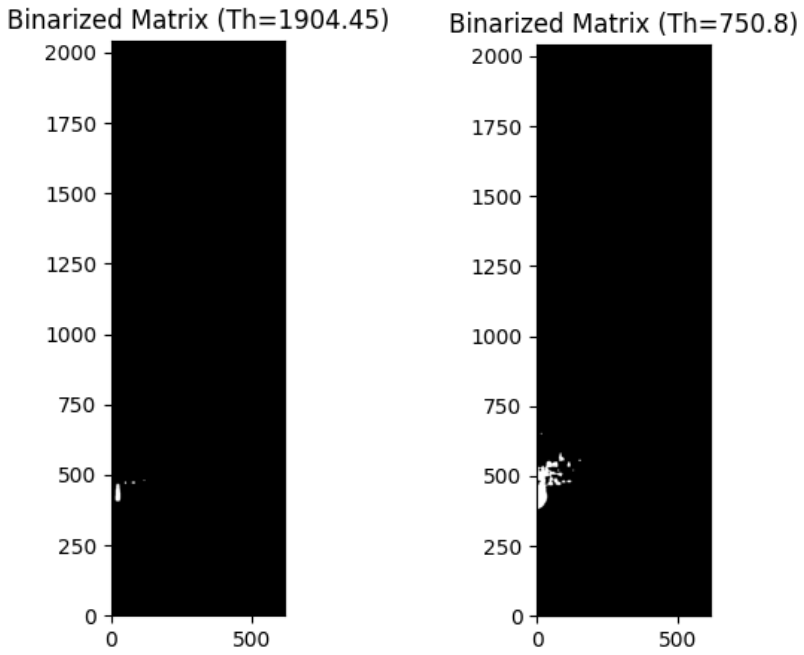


Figure 4.2: GC-IMS chromatogram of MRS 0h after drift time alignment

4.1.2 Results of ROI selection

Putting the average intensity matrix into the terative adaptive thresholding algorithm allows me to calculate a threshold value of 1904.45. Then I build a binarized matrix with this threshold value. The values in the mean intensity matrix that were larger than 1904.45 were set to 1 and the values that were lower than 1904.45 were set to 0, as shown in Figure 4.3a. We can find that many features are filtered out, by comparing the binarized matrix to the original matrix Figure 3.4. This result does not reach our expectations. Therefore, in the case of using the algorithm for the whole intensity matrix, it will lead to a threshold that will be too high.

I found two main reasons for this problem by watching the histogram of the mean



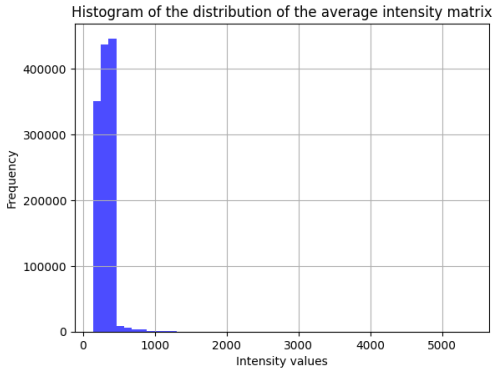
(a) Binarized Matrix (Th=1904.45) (b) Binarized Matrix(Th=750.8)

Figure 4.3: Binarized matrices

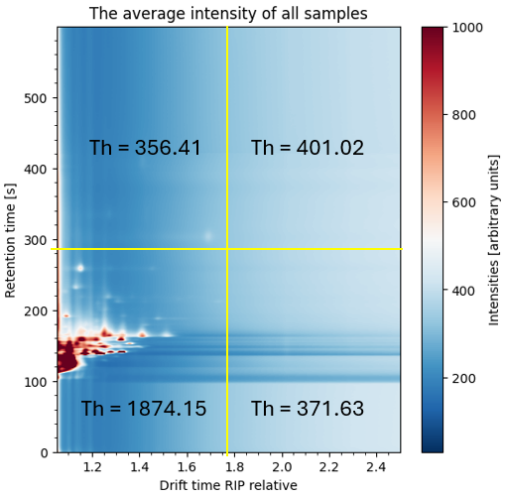
intensity matrix distribution, as shown in figure 4.4a. Firstly, the large difference between the maximum value (5393) and the minimum value (150) makes the initial threshold too high. This is influenced by the high values in the matrix. Secondly, the number of high values in the matrix is significantly less than the number of low values. This results in the number of values in region B (below the threshold) greatly exceeding the number of values in region A (above the threshold). Thus, limiting the threshold to decrease significantly during the iteration process.

To solve this problem . I divided the raw intensity matrix into four regions equally: upper left, upper right, lower left and lower right, as shown in figure 4.4b. Then i calculated the corresponding thresholds for each region. The thresholds for the upper-left, upper-right, and lower-right regions are 356.41, 401.02, and 371.63, respectively. The thresholds for these regions are close to the background values. Because there are no significant features in these regions. Since most of the features appear in the lower left region, the threshold value for this region is 1874.15. Finally, I calculated the average of the four thresholds as 750.81 and used this value to construct a binarized matrix. By comparing the features in figure4.3b and 3.4, we can see that the feature regions are approximately the same, so we can selecte 750.81 as the final threshold.

By recording the indices of the upper, lower, and right boundaries, I can map the corresponding times from the drift time and retention time arrays (as dts_end, rts_start, and rts_end). The average intensity matrix calculates these values as follows(dts_end=1.42, rts_start=111.72 s, rts_end=191.69 s). Then, I can cut our GC-IMS chromatogram



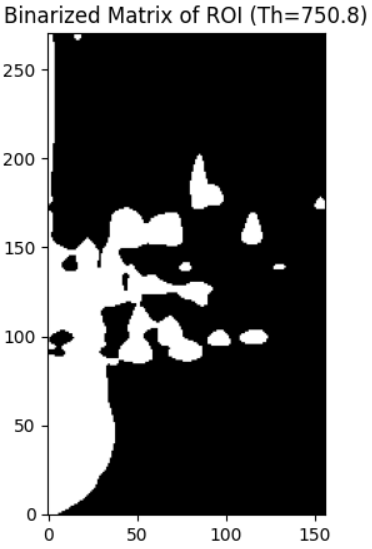
(a) Histogram of the distribution of the average intensity matrix



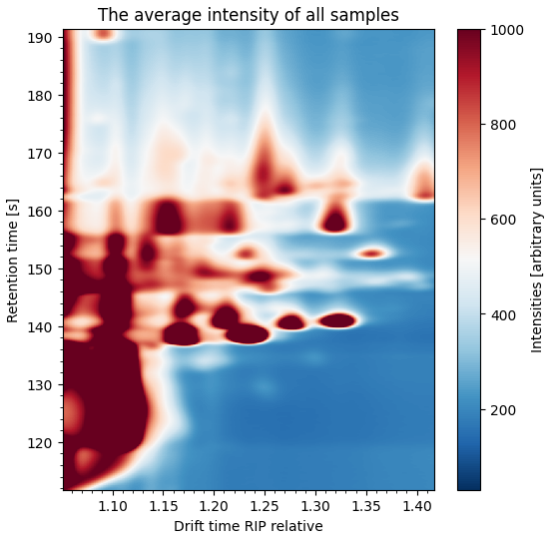
(b) GC-IMS chromatograms of the average intensity matrix with 4 regione

Figure 4.4: The process of average threshold analysis

based on these times. The figure4.5b shows the ROI region of the average intensity matrix. By comparing it with Figure3.4, we can see that the automatically selected ROI successfully extracted all the feature values, significantly reducing the matrix dimensions and background noise, which greatly facilitates subsequent data processing.



(a) Binarized matrix of ROI(Th=750.8)

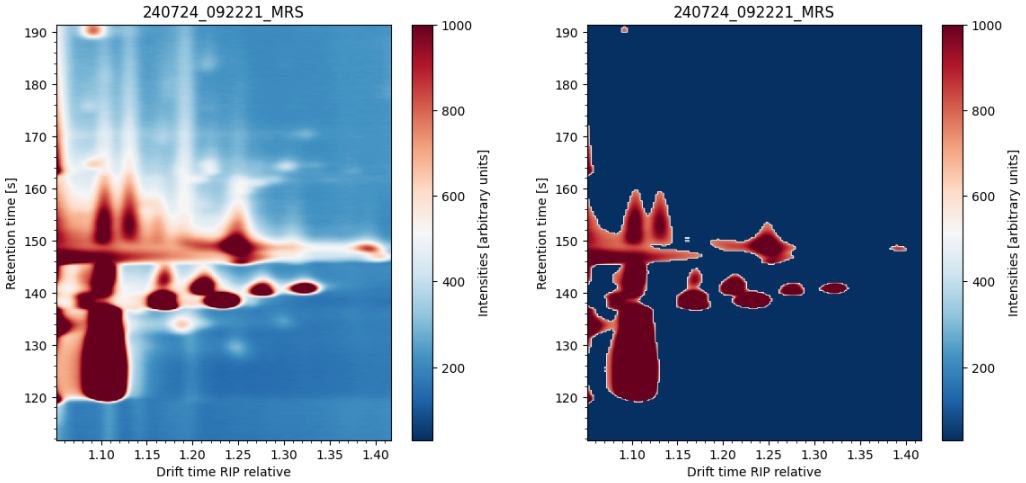


(b) ROI of GC-IMS chromatogram with averaged intensity matrices

Figure 4.5: Results for automatically selected regions of interest

4.1.3 Results of denoising and baseline correction

The result of denoising shown in the figure 4.6b. By observing Figure 4.6, there are a lot of white noise points in the left plot, while the white area of the right plot is almost completely removed leaving only the characteristic peaks.



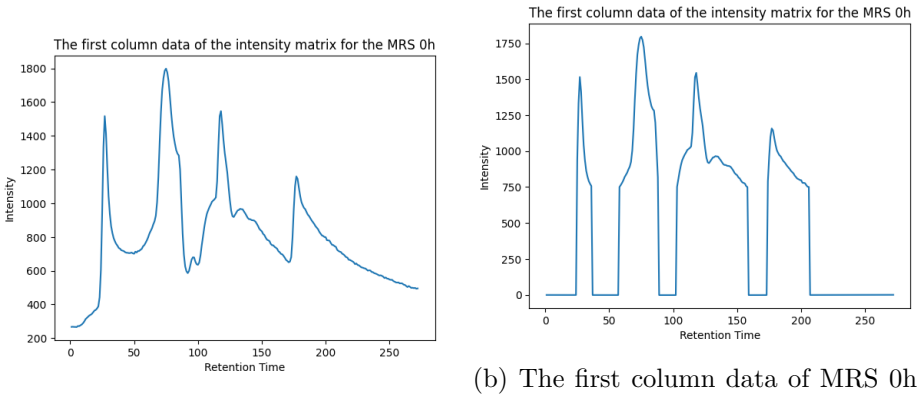
(a) GC-IMS chromatogram of MRS 0h after ROI selection

(b) GC-IMS chromatogram of MRS 0h after denoising

Figure 4.6: The process of MRS 0h denoising

In most cases, baseline drift is caused by fluctuations in retention time. This is because, in GC-IMS, retention time is directly related to the separation process of compounds in the gas chromatography column. During sample analysis, as retention time progresses, experimental conditions such as temperature, pressure, or flow rate may experience slight changes, leading to gradual shifts in the baseline signal. In contrast, drift time is mainly determined by the migration speed of molecules under the influence of an electric field, making it less susceptible to external interference. Therefore, drift time contributes minimally to baseline fluctuations.

To better observe the impact of baseline correction on the signal, I selected the first column of the intensity matrix, which corresponds to the intensity values varying with retention time. The left figure4.7a shows the original signal before baseline correction and denoising, where significant noise and baseline drift are evident. In contrast, the right figure4.7b shows the signal after baseline correction and noise removal. Through my many tests of different combinations of λ and p, I finally chose λ equal to 50000 and p equal to 0.01 to bring into the algorithm. In the corrected signal, the baseline has been effectively adjusted, noise has been significantly reduced, and the individual peaks are clearer, with the background intensity approaching zero. This process ensures that the extracted peaks in subsequent analyses are more reliable, reducing the interference of noise on the signal.



(a) The first column data of MRS 0h

(b) The first column data of MRS 0h after baseline correction

Figure 4.7: The process of MRS 0h baseline correction

4.1.4 Results of PCA score plot

The following conclusions can be reached from PCA score plots 4.8:

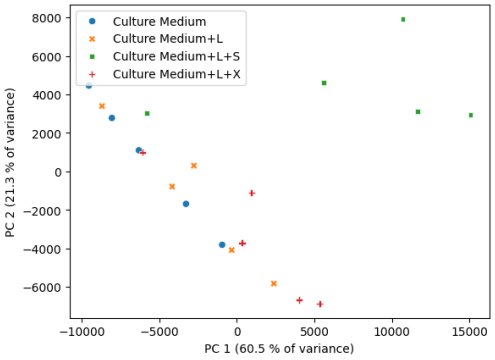
- In the PCA score plot of the original data (figure4.8a), PC1 explains 60.5% of the variance and PC2 explains 21.3% of the variance. This indicates that these two principal components effectively capture the main variation trends in the data. However, the distribution of data points is more dispersed, reflecting the variability among different categories. For example, “Culture Medium+L+S” (green squares) is far away from the other groups, while “Culture Medium+L” (orange crosses) and “Culture Medium+L+X” (red crosses) are clustered relatively close together. “Culture Medium” (blue dots) is more concentrated, but not clearly separated from the other categories.

- In the PCA score plot after pre-processing(figure4.8b), PC1 explains 48.9% of the variance and PC2 explains 17.1% of the variance. The decrease in variance explained by the principal components compared to the first plot suggests that the pre-processing may have removed some of the uncorrelated variables but retained the main trends in the data. The pre-processing resulted in a more centralized and compact distribution of data points. For instance, “Culture Medium” (blue) almost completely overlap, suggesting less internal variation in this sample category or effective removal of irrelevant noise. “Culture Medium+L” (orange) and “Culture Medium+L+X” (red) show more distinct distribution and are more clearly separated from the other categories. “Culture Medium+L+S” (green), although still slightly distant from the other groups, exhibits a more regular distribution compared to the unprocessed data.

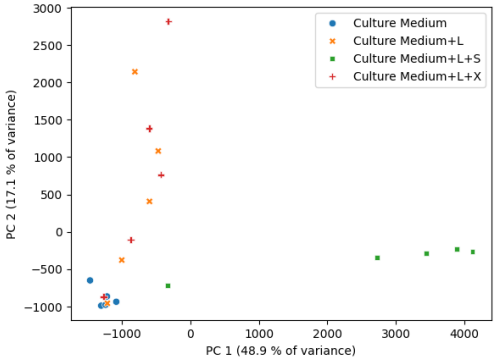
- Furthermore, in Figure4.8b, some sample points from the “Culture Medium+L” and “Culture Medium+L+X” groups are closely aligned with “Culture Medium” points, and one “Culture Medium+L+S” point is near the “Culture Medium” sample point. In Figure4.8c, these sample points correspond to the 0-hour experimental time point. This can be attributed to the fact that all Culture Medium

samples were in the same state before fermentation began, resulting in minimal differences.

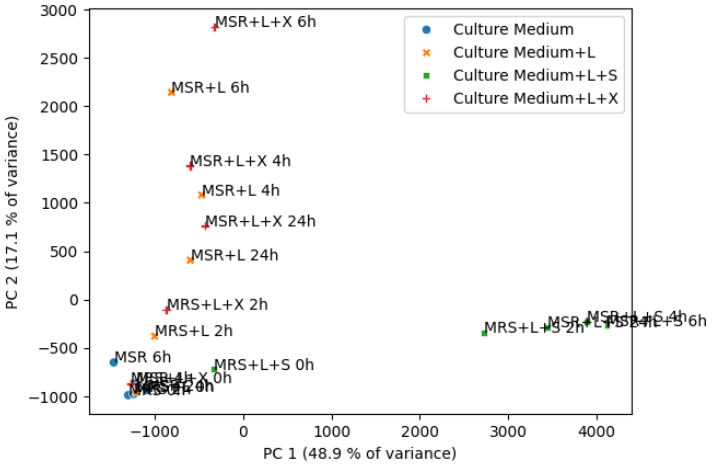
- Overall, the raw data has high variance, potentially influenced by noise or external factors, leading to a more dispersed distribution across categories. After pre-processing, noise and unrelated features are effectively removed, resulting in clearer clustering between sample groups. This highlights the importance of pre-processing in removing confounding factors from the data.



(a) PCA score plot of raw data



(b) PCA score plot after pre-processing



(c) Annotated PCA score plot after pre-processing

Figure 4.8: Comparison of PCA score plots

4.2 Results of Peak selection and machine uncertainty analysis

Thirdly, the PH technique was applied to select peaks on each sample. And the coffee data set was used to determine the uncertainty of the machine in the selection of peak locations.

4.2.1 Results of Peak selection

Results of PH and Peak table

The peak information detected by the PH algorithm is systematically recorded in a table(Example table 4.1, data from MRS 0h sample). This table provided the data necessary for the construction of the persistence plots. The table includes drift times, retention times, pixel coordinates, and birth and death levels for each peak, providing a comprehensive summary of the important features in the chromatogram.

peak number	compound	drift_time	x	ret_time	y	birth_level	death_level	score	peak	valley
1		1.104932	23	122.010	35	254.037423	0.000000	254.037423	True	False
2		1.107278	24	138.474	91	-254.027470	-0.000000	254.027470	False	True
3		1.231612	77	138.474	91	185.096175	34.063651	151.032524	True	False
4		1.165926	49	137.592	88	139.055231	34.078747	104.976484	True	False
5		1.323103	116	140.826	99	114.075486	34.060992	80.014494	True	False
...	...	...	...	...	...	...	...	...	...	...
8188		1.344216	125	166.110	185	-220.099555	-220.099526	0.000029	False	True
8189		1.402865	150	170.520	200	-220.093124	-220.093097	0.000026	False	True
8190		1.233958	78	180.516	234	-220.095925	-220.095905	0.000020	False	True
8191		1.292606	103	116.718	17	-220.078131	-220.078111	0.000020	False	True
8192		1.095548	19	115.542	13	35.091289	35.091274	0.000015	True	False

Table 4.1: Peak table of MRS 0h

Persistence plot

We used this table to build a persistence plot, as shown in figure 4.9 By observing this plot, we can find that the peak points are mainly distributed in the first and third quadrants that represent peaks and valleys in the data, respectively. A point in the first quadrant indicates that the feature is positive from a level of births and deaths, forming a more significant peak; while a point in the third quadrant with negative levels of births and deaths indicates the corresponding valley feature. In other words, PH can help us detect not only peaks but also valleys in our data. The distance from the diagonal line (where birth equals death) indicates the persistence of the feature. The farther the point is from the diagonal, the more significant the feature.

The first 20 significant peaks of GC-IMS chromatogram

In the analysis of the GC-IMS chromatogram (Figure 4.10a), the red areas represent regions that may contain peaks. By applying the PH algorithm, I repeatedly tested and filtered out the most important peaks, and ultimately settled on a peak value of 20. This is because at this value there is at least one peak detected in each red region of the GC-IMS chromatogram. These peaks are not only characterised by their persistence in the data, but they are also the most analytically significant parts.

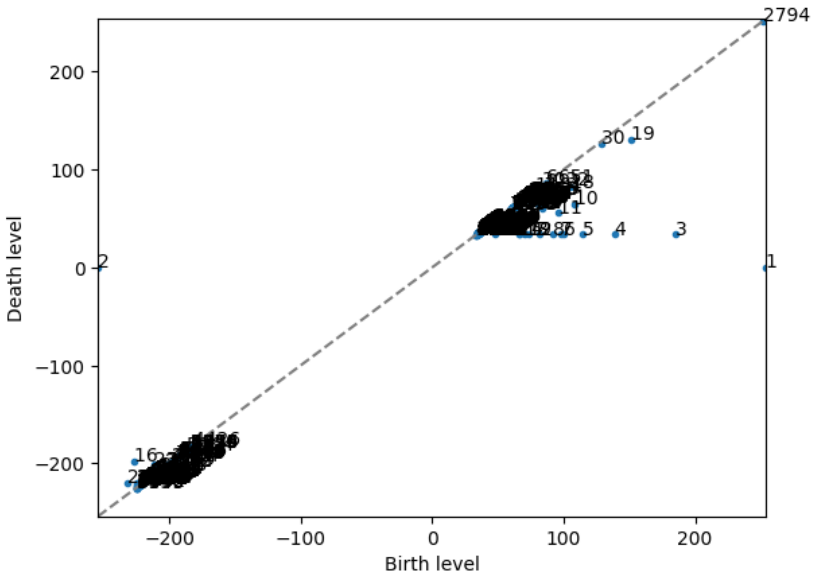
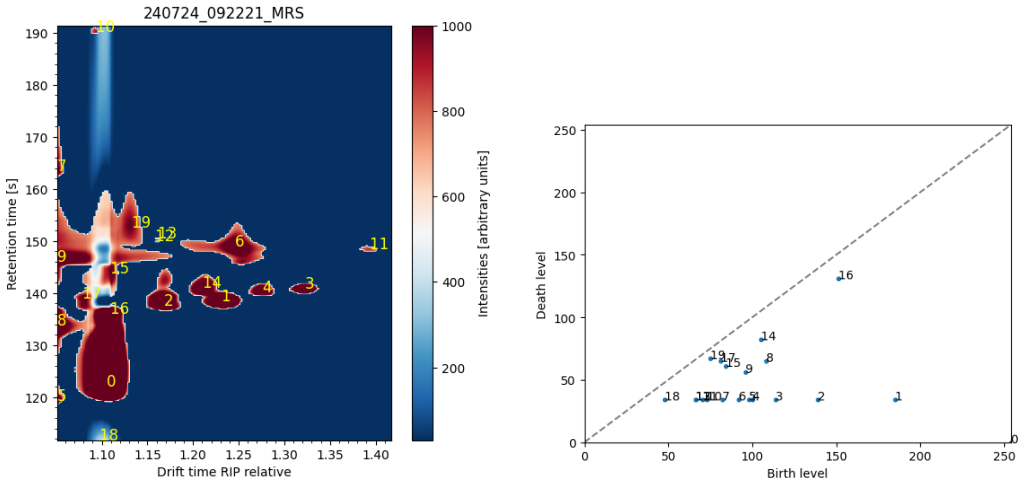


Figure 4.9: Persistence plot of MRS 0h sample

The table of peaks shows further details of each of the first 20 significant peaks (see table 4.2). Each row in the table represents a peak, providing the drift time, retention time, coordinate position in the matrix (x, y), and the birth and death levels for the peak. For example, peak 0 has the highest level of births (254.05) and 0 deaths, which means that it is the most persistent peak and is of extremely high importance. Whereas a peak like peak 19 has lower levels of births and deaths, indicating that it is relatively less stable. We can go ahead and construct the persistence graph based on this table. In the persistence plot (Figure 4.10b), these peaks are presented visually through the relationship between birth level and death level. We can observe that the death level of most of the peaks is on a horizontal line(death level= 34), meaning that these signalling features disappear at the same level. This is due to the fact that we have performed the baseline correction preprocessing which makes the baseline of these signals at the same level.

By combining the information in the icons above, we were able to determine that these 20 peaks were the most analytically significant portion of the sample. The information on drift times and retention times, as well as their persistence, helped us to get a more complete picture of the chemical composition of the sample. In subsequent analyses, we can focus on these most important peaks while filtering out unimportant or noisy data.

In order to observe how the first 20 significant peaks of each sample varied at different levels of fermentation in different media. I applied the PH algorithm to all the samples as shown in the figure4.11. There are four rows in the figure representing four different media which are MRS, MRS_L,MRS_L_S and MRS_L_X respectively. Each row has five columns corresponding to 0, 2, 4, 6, and 24 hours of sampling time. Each peak is



(a) GC-IMS chromatogram of MRS 0h and first 20 significant peaks

(b) Persistence plot of MRS 0h sample and first 20 significant peaks

Figure 4.10: MRS 0h sample and first 20 significant peaks

	compound	drift_time	x	ret_time	y	birth_level	death_level	score	peak	valley	peak_number
0		1.104932	23	122.010	35	254.052104	0.000000	254.052104	True	False	1
1		1.231612	77	138.474	91	185.053552	34.065928	150.987624	True	False	2
2		1.168272	50	137.592	88	139.046623	34.073152	104.973472	True	False	3
3		1.323103	116	140.826	99	114.013520	34.059204	79.954315	True	False	4
4		1.276185	96	140.238	97	100.030128	34.076783	65.953345	True	False	5
5		1.050975	0	119.364	26	98.084744	34.064792	64.019952	True	False	6
6		1.245687	83	149.058	127	92.025662	34.077248	57.948415	True	False	7
7		1.050975	0	163.464	176	82.020812	34.059678	47.961134	True	False	8
8		1.050975	0	133.770	75	108.060284	65.051634	43.008651	True	False	9
9		1.050975	0	146.118	117	96.056205	56.089865	39.966339	True	False	10
10		1.093202	18	190.218	267	73.057076	34.067510	38.989565	True	False	11
11		1.393481	146	148.470	125	70.096439	34.062336	36.034103	True	False	12
12		1.158888	46	149.940	130	66.095261	34.065945	32.029317	True	False	13
13		1.161234	47	150.528	132	66.094344	34.069842	32.024501	True	False	14
14		1.210499	68	141.120	100	105.081371	82.021838	23.059533	True	False	15
15		1.109624	25	143.766	109	84.059057	61.027185	23.031872	True	False	16
16		1.109624	25	136.122	83	151.069824	131.004340	20.065484	True	False	17
17		1.079127	12	139.062	93	81.060798	65.069541	15.991257	True	False	18
18		1.097894	20	111.720	0	48.063841	34.058402	14.005438	True	False	19
19		1.133083	35	152.586	139	75.081790	67.028359	8.053431	True	False	20

Table 4.2: Peak table of the first 20 significant peaks of MRS 0h

represented by a small yellow square.

In the MRS medium, there were no significant new peaks generated with time changes because no microorganisms were added. In MRS_L_S medium. at 0 hours, it can be seen that the areas of the detected peaks are all smaller. In the 2-4 hour samples, it can be seen that the area of the previous peaks has become larger and several new peaks have appeared. In the 6-24 hour samples, some of the peaks that were previously detected as

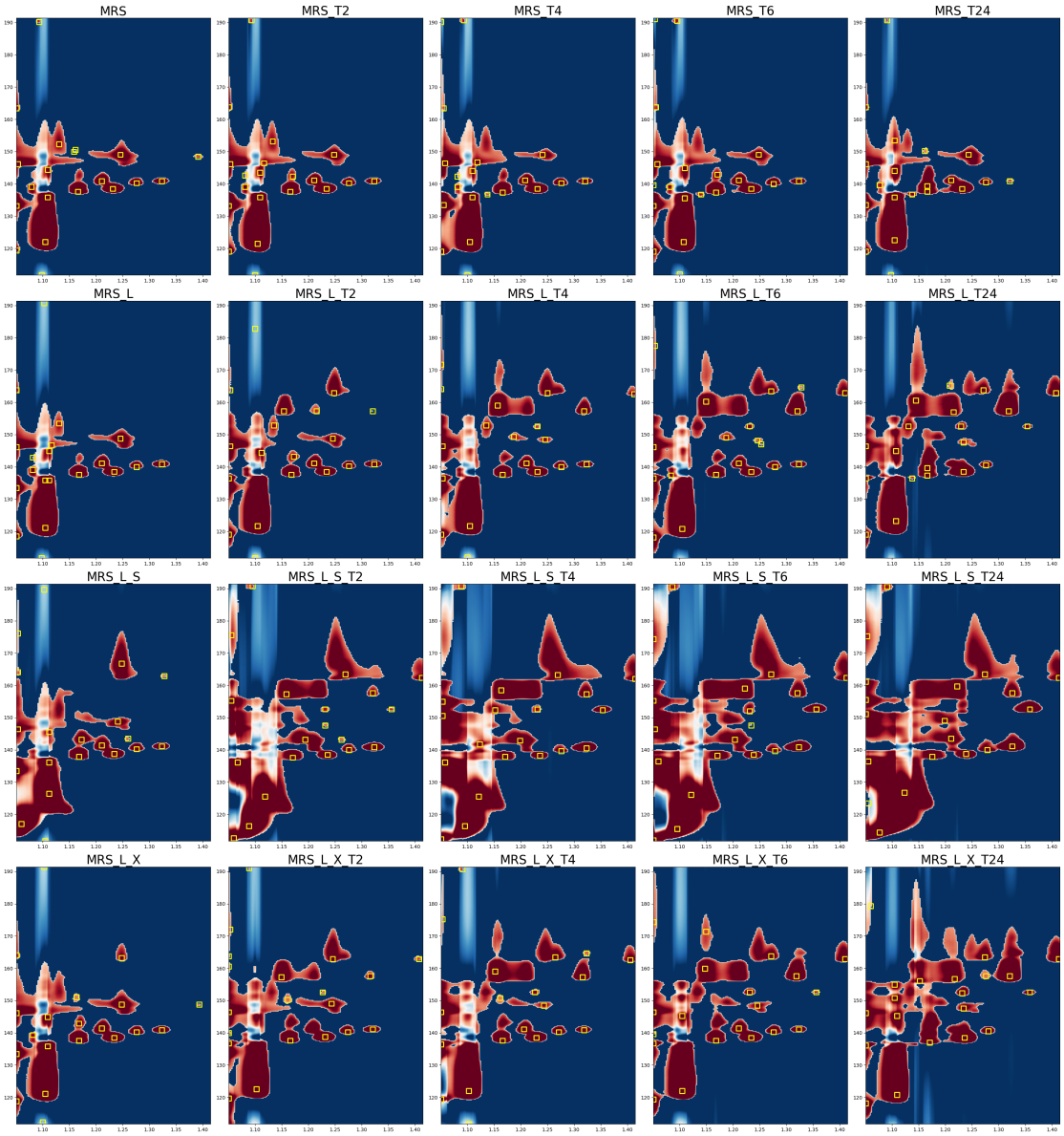


Figure 4.11: First 20 significant peaks in each sample

significant diminished in intensity over time. And some new peaks increased in intensity, replacing the previous peaks in the top 20 significant peaks. Since MRS_L_X is a control, we can compare it to MRS_L together. The two groups of samples present a similar peak distribution in the initial phase (T0). With time, the MRS_L_X group has some peaks with more intensity than the MRS_L group in the later time (T6 and T24), and some new peaks appear. This also makes the first 20 significant peaks in the sample not exactly in the same position.

Overall the PH technique can successfully help us to identify the peaks in the samples and helps us to observe the changes in the peaks in the samples. But it is very inefficient

to compare the peaks in the samples manually so we next used machine learning to identify the significant peaks in the dataset.

4.2.2 Result of machine uncertainty analysis

By using the standard deviation formula 3.21, the final results are summarized in the table 4.3. The last four columns of the table 4.4 show the standard deviation for each peak, including the deviations in the x and y coordinates, drift time, and retention time. We can observe that the results obtained are very small . It shows that the position of the peak does not change much after similar samples have been repeatedly tested by the machine. Depending on the type of sample (roasted coffee, which does not change its properties during the measurement), the variation of the peaks provides the instrumental uncertainty (no change in the sample itself). Therefore, the uncertainty we calculated is the uncertainty of the machine (HS-GC-IMS) during various measurements. With this data we can conclude that the machine we are using performs very consistently.

	base_x	base_y	base_drift_time	base_ret_time	sample_1_x	sample_1_y	sample_1_drift_time	sample_1_ret_time	sample_2_x	sample_2_y	...	sample_8_drift_time	sample_8_ret_time	sample_9_x	sample_9_y	sample_9_drift_time	sample_9_ret_time
0	22	420	1.102104	124.656	22.0	409.0	1.102104	121.422	22	413	...	1.099769	123.480	21	420	1.099769	124.656
1	77	468	1.230527	138.768	78.0	467.0	1.232862	138.474	78	468	...	1.228192	138.768	76	468	1.228192	138.768
2	78	564	1.232862	166.992	79.0	563.0	1.235197	166.698	78	563	...	1.230527	166.698	77	563	1.230527	166.698
3	39	441	1.141798	130.830	39.0	440.0	1.141798	130.536	39	441	...	1.139463	130.830	38	441	1.139463	130.830
4	95	550	1.272556	162.876	95.0	550.0	1.272556	162.876	95	550	...	1.272556	162.876	95	550	1.272556	162.876
5	113	654	1.314585	193.452	114.0	655.0	1.316920	193.746	113	654	...	1.312250	193.746	114	654	1.316920	193.452
6	40	503	1.144133	149.058	NaN	NaN	NaN	NaN	40	504	...	1.141798	149.058	40	503	1.144133	149.058
7	115	533	1.319255	157.878	115.0	534.0	1.319255	158.172	115	533	...	1.316920	157.878	114	533	1.316920	157.878
8	96	474	1.274891	140.532	96.0	474.0	1.274891	140.532	96	474	...	1.272556	140.532	95	474	1.272556	140.532
9	50	463	1.167483	137.298	49.0	463.0	1.165148	137.298	50	463	...	1.165148	137.298	49	463	1.165148	137.298

Table 4.3: Position of the first 10 peaks in all samples

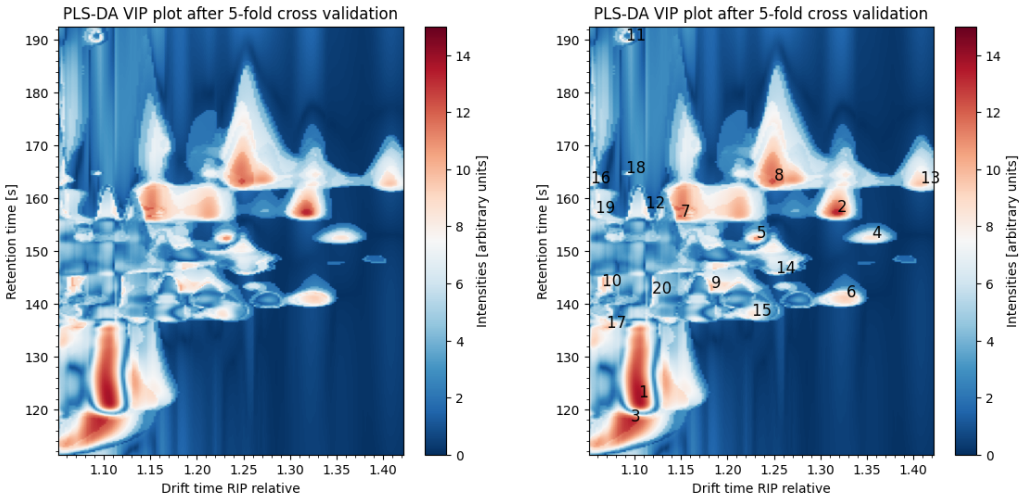
x_incertezza	y_incertezza	drift_time_incertezza	ret_time_incertezza
0.421637	3.765339	0.000985	1.107010e+00
0.823273	0.421637	0.001922	1.239613e-01
0.567646	0.516398	0.001325	1.518209e-01
0.516398	0.516398	0.001206	1.518209e-01
0.421637	0.000000	0.000985	0.000000e+00
0.567646	0.567646	0.001325	1.668880e-01
0.916125	0.353553	0.002139	1.039447e-01
0.421637	0.316228	0.000985	9.297096e-02
0.567646	0.000000	0.001325	2.995911e-14
0.516398	0.000000	0.001206	0.000000e+00

Table 4.4: Results of the standard deviation of the positions of the first 10 samples

4.3 Results with significant peaks in VIP scores and their visualisation

Finally, the VIP score and PH were combined to select the important peaks and confirm their drift and retention times. Areas of these times were selected in all samples to generate visualisations for observing the variation of significant peaks in different samples.

I converted the VIP scoring data into the GC-IMS chemical mapping phase data format and visualised it as shown in Figure 4.12a. With the added drift time and retention time, the location of each characteristic peak can be more accurately located. By adding drift times and retention times, it was possible to more precisely locate the position of each feature peak. The red areas in the figure represent feature peaks with higher scores. These feature peaks are more capable of distinguishing between different samples, which indicates that they appear to be more important in the model with high contribution. This visualisation helps us to visually identify which peaks play a key role in sample classification.



(a) VIP plot after 5-fold cross validation (b) The first 20 peaks of VIP plot

Figure 4.12: VIP plots

Subsequently, We then filtered and ranked these peaks using the PH technique. The results are shown in Table 4.5. In this table, we can see detailed information about each peak (drift time, retention time, coordinates in the picture, normalised birth and death levels, respectively). The peak number is sorted according to the score of the peak, in other words according to the importance of the peak. Peak 1, whose position is at drift time 1.1 retention time 122.3s, has a score of 245.08. It means that he is the most important peak in our whole dataset.

Then, I marked the selected peaks on the VIP graph by using the drift times and retention times provided in the table. As shown in the figure 4.12b. To ensure that each high score region in the graph contains at least one selected peak. Through several experiments, I eventually determined that 20 peaks were chosen as the optimal number to analyse. If more than 20 peaks are selected, the same region may contain multiple peaks, resulting in redundant information; However, if less than 20 peaks are selected, some important high scoring regions may be missed. Therefore, the selection of 20 peaks ensures both the brevity of the data and that all key regions are effectively characterised.

peak number	compound	drift_time	x	ret_time	y	birth_level	death_level	score	peak	valley
1		1.104932	23	122.304	37	254.086283	0.000000	254.086283	True	False
2		1.318411	114	157.584	157	244.045879	82.045803	162.000076	True	False
3		1.095548	19	117.894	22	245.059806	93.084257	151.975548	True	False
4		1.355946	130	152.586	140	159.091645	17.076547	142.015098	True	False
5		1.231612	77	152.586	140	200.087355	60.094763	139.992593	True	False
6		1.327795	118	141.414	102	171.088926	40.000023	131.088903	True	False
7		1.149504	42	156.702	154	228.039243	98.003616	130.035626	True	False
8		1.250379	85	163.464	177	232.006930	105.022536	126.984394	True	False
9		1.182347	56	143.178	108	190.028809	76.066524	113.962285	True	False
10		1.065051	6	143.472	109	188.046842	77.061332	110.985510	True	False
11		1.090856	17	189.924	267	161.023480	52.047461	108.976019	True	False
12		1.111970	26	158.172	159	167.075333	67.074246	100.001088	True	False
13		1.407556	152	162.876	175	177.074082	78.058399	99.015683	True	False
14		1.252725	86	145.824	117	161.024014	62.047812	98.976202	True	False
15		1.226920	75	137.886	90	157.038429	65.090259	91.948170	True	False
16		1.053321	1	162.876	175	163.044441	75.086455	87.957986	True	False
17		1.069743	8	135.534	82	188.005292	102.089689	85.915603	True	False
18		1.090856	17	164.934	182	123.010963	47.078726	75.932237	True	False
19		1.058013	3	157.290	156	159.043027	85.046220	73.996807	True	False
20		1.119007	29	142.002	104	171.065993	99.092892	71.973101	True	False

Table 4.5: Peak table of VIP Scores

To better observe how these 20 peaks change in each sample, I used the following approach. Firstly, the drift time and retention time of each peak in the table 4.5 was used as a centre point. Then I chose a small range interval around this centre point. Specifically, I set the drift time range for each peak to plus or minus 0.03, while the retention time range was set to plus or minus 5 seconds. These two parameters together construct a small rectangular region in which we can more easily visualise the shape of the peaks. These two parameters are the best results after I have adjusted the parameters many times. Through repeated adjustments, I found that this range not only captures the full picture of the peaks clearly, but also effectively reduces other irrelevant peaks from entering the region and avoiding them to interfere with the judgement. Next, I applied this region to all 20 peaks. These small rectangular regions are then further applied to all samples in turn. In this way, we can capture the shapes of these 20 peaks in each sample and be able to compare how they vary across samples. Finally, I have summarised the regions of the top 20 peaks from all the samples, as shown in figure 4.13. The purpose of this is to show more clearly how the individual peaks change as the fermentation time increases. Additionally it is possible to observe how the same peaks behave in different media.

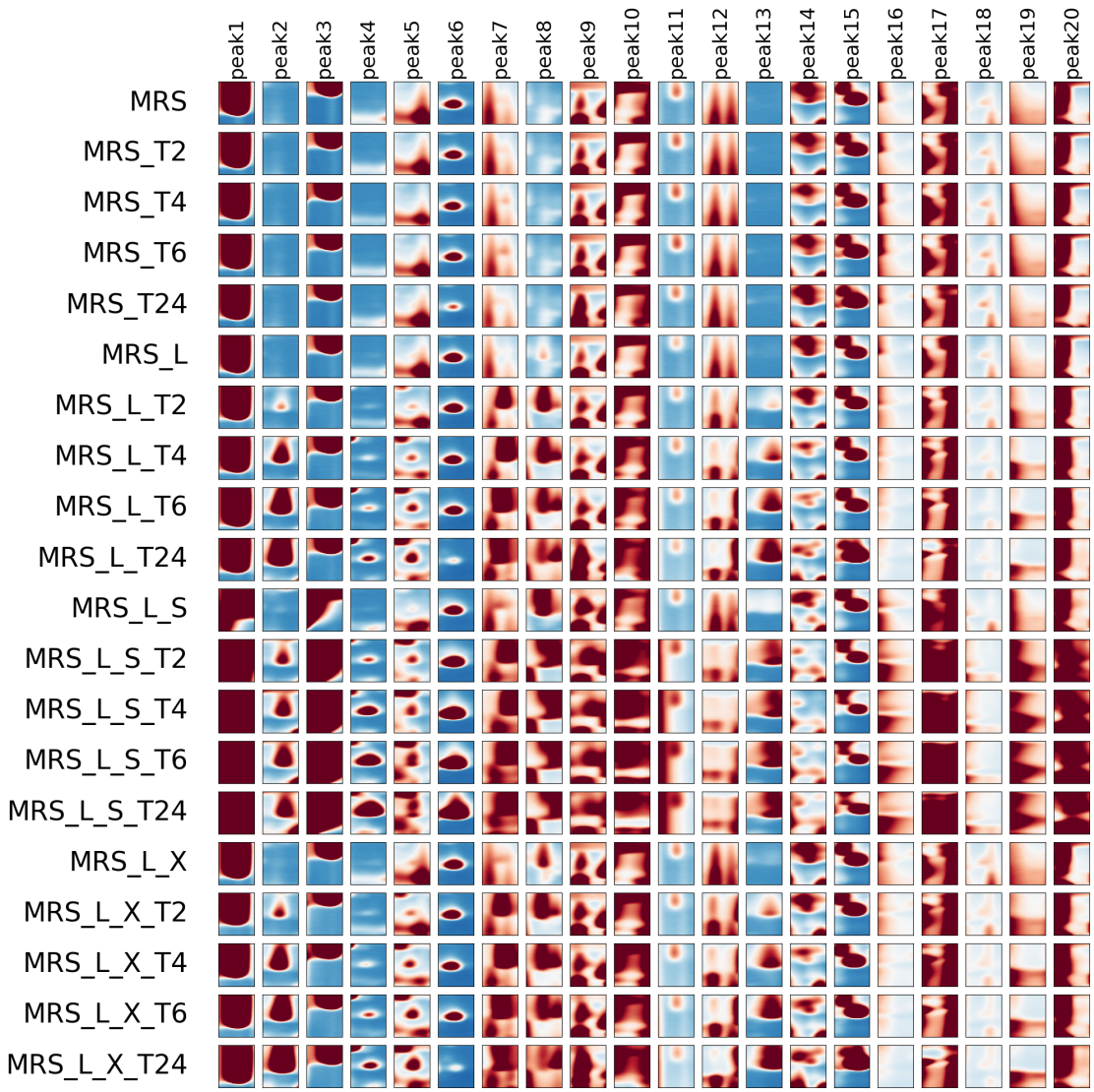


Figure 4.13: The first 20 peaks of each sample

In this figure, each row represents a sample and each column represents a significant peak. We can get the following conclusions by observing this this figure:

1. We can see that all the samples with T=0 h (codes MRS, MRS_L, MRS_L_S, and MRS_L_X respectively) have very similar areas and colours of their peaks. This is because these samples were sampled before the fermentation started and the particular VOCs had not been produced in the medium.

2. we can also see that peak 1 maintains a very high area and dark colour in all the samples. It means that this VOC will not change with the medium and time, it is a stable peak.

3. Peak 2, Peak 4, peak 8 and peak 13 are not present in the MRS medium or have

very low intensity. However, in MRS_L, MRS_L_S, and MRS_L_X medium, the area and intensity of these three peaks increase as the fermentation time passes. Among them, the changes of peak 2 peak 8 and peak 13 are the most obvious. This means that these VOCs are metabolites of microorganisms, which can be used as one of the bases for us to judge the degree of fermentation.

4. Peak 3, Peak 4, Peak10, Peak 17, Peak 19, Peak 20 grew slowly during the fermentation of MRS_L and MRS_L_X. But these peaks were highly variable during the fermentation of MRS_L_S. This indicates that these VOCs may be metabolites of brewer's yeasts. So the intensity of these VOCs will be much higher than the other groups as time increases.

5. There are also some peaks that are very intense and large at MRS and T = 0 h, but with the addition of microorganisms and fermentation, they become less intense and smaller than before. For example peak 6 and peak 14. This is because this type of VOCs may be consumed during the fermentation process and their intensity slowly becomes lower as the fermentation progresses. However, in Peak 6, when brewer's yeast was added to the medium, the intensity and area of this peak increased instead. This situation may be due to the fact that, this VOC is one of the metabolites of brewer's yeast. But it is not the case for Peak 6, when brewer's yeast was added to the medium, the intensity and area of this peak increased instead. This situation could be because, this VOC is one of the metabolites of brewer's yeast.

Chapter 5

Conclusion

5.1 Research summary

With the increasing popularity of fermented foods, automation and intelligent monitoring of the fermentation process has become one of the popular research issues.

This thesis focuses on the potential application of machine learning and topological data analysis for solving this problem and discusses its results.This study is based on sampling data from four different sets of media at different fermentation time points and evaluates the results of the data after pre-processing using PH and PLS-DA to select peaks.The variation of peaks over the duration of fermentation was analysed by constructing peak sample plots. The results show that we can see in Figure 4.13 how the selected peaks changed over time in different media. Among them, peak 2, peak 3 and peak 13 showed the most obvious changes.

This study represents the possibility of monitoring fermentation by selecting characteristic peaks. It indicated that this technique will have the potential to be applied in the production industry in the future.However it is necessary to validate it by more data to confirm its stability.

5.2 Limited discussion

This study automated most of the steps in the GC-IMS data process. However, there are still some parameters that need to be manually adjusted (e.g., parameters for baseline correction) that cannot be selected automatically by the data. And the number of samples was limited to do a more in-depth study. Therefore, the number of samples should be expanded in the future so that there is more data to train the model to achieve full automation of this technique. This will reduce the manipulation of researchers and operators and results can be obtained from a more objective point of view.

5.3 Future research directions

Future research could be done by identifying the specific chemical composition of important VOCs. Then observe how these VOCs change by changing the fermentation environmental parameters. This will lead to the control of fermentation.

And we can incorporate advanced machine learning models such as deep learning or ensemble methods. The selected peaks will be classified to know which VOCs are positive feedbacks for fermentation and which are unfavourable for fermentation to proceed.

Bibliography

[1] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

[2] Joscha Christmann, Sascha Rohn, and Philipp Weller. gc-ims-tools a new python package for chemometric analysis of gc–ims data. *Food Chemistry*, 394:133476, 2022. ISSN 0308-8146. doi: https://doi.org/10.1016/j.foodchem.2022.133476. URL https://www.sciencedirect.com/science/article/pii/S0308814622014388.

[3] GA Eiceman, EG Nazarov, JE Rodriguez, and JF Bergloff. Positive reactant ion chemistry for analytical, high temperature ion mobility spectrometry (ims): Effects of electric field of the drift tube and moisture, temperature, and flow of the drift gas. *Int. J. Ion Mobil. Spectrom*, 1:28–37, 1998.

[4] Mireia Farrés, Stefan Platikanov, Stefan Tsakovski, and Romà Tauler. Comparison of the variable importance in projection (vip) and of the selectivity ratio (sr) methods for variable selection and interpretation. *Journal of Chemometrics*, 29(10): 528–536, 2015.

[5] Mario Fordellone, Andrea Bellincontro, and Fabio Mencarelli. Partial least squares discriminant analysis: A dimensionality reduction method to classify hyperspectral data. *arXiv preprint arXiv:1806.09347*, 2018.

[6] Valérie Gabelica and Erik Marklund. Fundamentals of ion mobility spectrometry. *Current opinion in chemical biology*, 42:51–59, 2018.

[7] Chang Gao, Rui Wang, Fang Zhang, Zhengchen Sun, and Xianghong Meng. The process monitors of probiotic fermented sour cherry juice based on the hs-gc-ims. *Microchemical Journal*, 180:107537, 2022.

[8] D Vaughan Griffiths and Ian Moffat Smith. *Numerical methods for engineers*. Chapman and Hall/CRC, 2006.

[9] Shuang Gu, Jing Zhang, Jun Wang, Xiangyang Wang, and Dongdong Du. Recent development of hs-gc-ims technology in rapid and non-destructive detection of quality and contamination in agri-food products. *TrAC Trends in Analytical Chemistry*, 144:116435, 2021. ISSN 0165-9936. doi: https://doi.org/10.1016/j.trac.2021.116435. URL https://www.sciencedirect.com/science/article/pii/S0165993621002582.

[10] MN Hasan, MZ Sultan, and M Mar-E-Um. Significance of fermented food in nutrition and food science. *Journal of Scientific Research*, 6(2):373–386, 2014.

[11] Stefan Huber. Persistent homology in data science. In *Data Science–Analytics and Applications: Proceedings of the 3rd International Data Science Conference–iDSC2020*, pages 81–88. Springer, 2021.

[12] Melanie Jünger, Bertram Bödeker, and Jörg Ingo Baumbach. Peak assignment in multi-capillary column–ion mobility spectrometry using comparative studies with gas chromatography–mass spectrometry for voc analysis. *Analytical and bioanalytical chemistry*, 396:471–482, 2010.

[13] Bruno Kolb and Leslie S Ettre. *Static headspace-gas chromatography: theory and practice*. John Wiley & Sons, 2006.

[14] Federico Marini. *Chemometrics in food chemistry*. Newnes, 2013.

[15] Andrew Nailman. Pca: An unsupervised dimensionality reduction technique, 2023. URL https://machinelearningmodels.org/pca-an-unsupervised-dimensionality-reduction-technique/.

[16] Sergio Oller-Moreno, Antonio Pardo, Juan Manuel Jiménez-Soto, Josep Samitier, and Santiago Marco. Adaptive asymmetric least squares baseline estimation for analytical instruments. In *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, pages 1–5. IEEE, 2014.

[17] Hadi Parastar, Joscha Christmann, and Philipp Weller. Automated 2d peak detection in gas chromatography-ion mobility spectrometry through persistent homology. *Analytica Chimica Acta*, 1289:342204, 2024. ISSN 0003-2670. doi: https://doi.org/10.1016/j.aca.2024.342204. URL https://www.sciencedirect.com/science/article/pii/S0003267024000059.

[18] PerfectLight. The basic principles and structure of gas chromatography instrument. https://https://www.perfectlight.com.cn/news/detail-99.html, 2022.

[19] Ewa Szymańska, Antony N Davies, and Lutgarde MC Buydens. Chemometrics for ion mobility spectrometry data: recent advances and future prospects. *Analyst*, 141 (20):5689–5708, 2016.

[20] Shaw Talbi. Persistent homology | introduction python example code. URL https://www.youtube.com/watch?v=5ezFcy9CIWE.