

Detection and integration of chromatographic peaks using theoretical peak fitting

Juho Kuikka

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 24.05.2022

Supervisor

Prof. Juho Rousu

Advisor

Dr Thiago Brito



Aalto University
School of Science

Copyright © 2022 Juho Kuikka

Author Juho Kuikka

Title Detection and integration of chromatographic peaks using theoretical peak fitting

Degree programme Computer, Communication and Information Sciences

Major Machine Learning, Data Science and Artificial Intelligence **Code of major** SCI3044

Supervisor Prof. Juho Rousu

Advisor Dr Thiago Brito

Date 24.05.2022**Number of pages** 56**Language** English

Abstract

Peak detection is a fundamental part of chromatography data analysis. In liquid chromatography-mass spectrometry (LC-MS) method development, accurate peak detection is crucial to producing reliable results. Noise in the chromatogram, among other things, can make peak detection challenging. In this Master's thesis, a literature review of peak detection and integration methods is presented, as well as methods used to estimate noise in the chromatogram signal prior to peak detection. A novel framework for detecting chromatographic peaks that uses theoretical peak shapes is also introduced.

The second part of this thesis consists of a case study where the goal is to reduce the relative standard deviation of the integrated peak areas, since this parameter correlates with the method performance. A peak detection and integration framework was developed and applied to data sets produced with the LC/MS method. The framework had a couple of options for estimating noise in the chromatogram data and detecting the start and end points of the peak. The methods for detecting the start and end points of peaks is based on fitting a theoretical peak to the chromatograms. The relative standard deviations obtained with the framework are compared to the relative standard deviations obtained with a known peak detection method from the literature and to the currently used method in the analyzers.

The proposed framework performed well compared to the other methods, despite the noise in the data and the varying peak shapes. In most of the analysed data sets, the proposed framework was able to produce lower relative standard deviations of the integrated areas. It was concluded that fitting a theoretical peak improved the precision of the detection and integration of chromatographic peaks.

Keywords Peak detection, Peak fitting, LC-MS, Relative standard deviation

Tekijä Juho Kuikka

Työn nimi Kromatografia piikkien tunnistus ja integrointi teoreettisen piikkin sovituksen avulla

Koulutusohjelma Computer, Communication and Information Sciences

Pääaine Machine Learning, Data Science and Artificial Intelligence **Pääaineen koodi** SCI3044

Työn valvoja Prof. Juho Rousu

Työn ohjaaja Dr Thiago Brito

Päivämäärä 24.05.2022**Sivumäärä** 56**Kieli** Englanti

Tiivistelmä

Piikin tunnistus on oleellinen osa kromatografialla mitatun datan analysointia. Tarkka piikin tunnistus on hyvin tärkeä osa nestekromatografiaan ja massaspektrometri-
aan (liquid chromatography-mass spectrometry, LC-MS) perustuvien menetelmien kehityksessä, sillä tarkka piikin tunnistus mahdollistaa luotettavat analyysitulokset. Kohina kromatogrammeissa, muiden mahdollisten häiriöiden lisäksi, voi aiheuttaa haasteita piikin tunnistuksessa. Tässä diplomityössä esitellään sekä kirjallisuuskatsauksella piikin tunnistus ja integrointi menetelmiä, että menetelmiä kohinan arviointiin ennen piikin tunnistusta. Työssä esitellään myös uusi viitekehys kromatografialla mitatun datan piikin tunnistukseen ja integrointiin.

Tämän diplomityön toinen osa koostuu tapaustutkimuksesta, jossa tavoitteena oli pienentää integroitujen piikkien pinta-alojen suhteellista keskihajontaa. Tämä parametri korreloi metodin suorituskyvyn kanssa. Viitekehys piikin tunnistukseen ja integrointiin kehitettiin käyttäen LC-MS:llä tuotettua mittausdatajoukkoa. Viitekehys sisälsi vaihtoehtoja sekä kohinatason laskemiseen, että piikkien integroinnin alku- ja loppupisteiden arviointiin. Piikkien integroinnin alku- ja loppupisteiden arviointi perustuu teoreettisen piikinmuodon sovitukseen. Viitekehysten menetelmiä käyttäen saavutettuja piikin pinta-alan suhteellisia keskihajontoja vertaillaan sekä vastaaviin kirjallisuuservoihin että analysaattorin nykyisen menetelmän tuottamiin arvoihin.

Ehdotettu viitekehys suoriutui paremmin kuin verrokkit siitä huolimatta, että data oli kohinaista ja sisälsi vaihtelevia piikkien muotoja. Ehdotetut menetelmät viitekeh-
yksessä onnistuivat tuottamaan pienempiä suhteellisia keskihajontoja integroitujen piikkien pinta-aloissa useimmissa analysoiduissa datajoukoissa. Tulokset osoittivat, että teoreettisen piikinmuodon sovittaminen paransi kromatogrammien analysoinnin tarkkuutta.

Avainsanat Piikin tunnistus, Piikin sovitus, LC-MS, suhteellinen keskihajonta

Preface

I want to thank Professor Juho Rousu and Dr Thiago Brito for discussions and feedback on the thesis

Espoo, 24.5.2022

Juho Kuikka

Contents

Abstract	3
Abstract (in Finnish)	4
Preface	5
Contents	6
Symbols and abbreviations	8
1 Introduction	9
2 Liquid chromatography - mass spectrometry	11
2.1 Liquid chromatography	11
2.2 Mass spectrometry	12
2.3 Data processing	15
2.4 Quality metrics	16
3 Peak detection and fitting algorithms	18
3.1 Open-source peak detection algorithms	18
3.2 Continuous wavelet transform	20
3.3 Fitting a theoretical peak	22
3.4 Baseline estimation	24
3.5 Nonlinear least-squares fitting	26
4 Framework	29
4.1 Motivation	29
4.2 Software	30
4.3 Data acquisition	30
4.3.1 File processing	30
4.3.2 Reorganizing the collected data	31
4.4 Peak detection and integration pipeline	32
4.4.1 Data processing	32
4.4.2 Peak integration function	33
4.4.3 Integrating peaks based on fitted theoretical peaks	34
4.4.4 Adding baseline and CWT method	36
4.5 Execution of the methods	37
5 Experiments and results	39
5.1 Integration of fitted and real peak	39
5.2 Channel and baseline effect to results	41
5.3 Comparison of the methods	44
6 Conclusion	51

References

Symbols and abbreviations

Symbols

Operators

Abbreviations

LC-MS	Liquid chromatography–mass spectrometry
m/z	mass-to-charge-ratio
s/n	signal-to-noise ratio
QC	Quality control
RSD	Relative standard deviation
CWT	Continuous wavelet transform
ESI	electrospray ionisation source
APPI	atmospheric pressure photo-ionisation
APCI	atmospheric pressure chemical ionisation
TOF	time-of-flight
MAD	Median absolute deviation
EMG	Exponentially Modified Gaussian
GMG	half Gaussian modified Gaussian
PLMG	parabolic-Lorentzian
AGD	asymmetric Gaussian distribution
GEX	Generalized exponential
PMG	polynomial modified Gaussian

1 Introduction

The goal of the project is to improve signal peak detection and integration of chromatography data produced with liquid chromatography–mass spectrometry method (LC-MS). LC-MS technique is widely used technique for analyzing small molecule samples such as metabolites from plasma, blood, serum, urine, and tissue [1]. LC-MS method uses separation ability of liquid chromatography which is based on chemical properties of the compounds in a sample [2]. After the compounds of a sample are separated in the LC-phase, the compound selected for measurement travel to the mass spectrometer where the molecules of the compound are separated based on mass-to-charge-ratio (m/z). The m/z ratio is calculated by dividing the mass of the molecule by its electric charge. There are different ionization methods for separating molecules such as electrospray ionization and chemical ionization [3]. A molecule desired to be measured from the compound with a specific m/z ratio is measured with a mass spectrometer. This measurement forms a signal. This signal usually includes a peak from which its intensity can be measured. The integrated peak intensity can then be used in result calculations. The problem in the peak detection and integration is that the signal and the peak include noise [4]. The signal always include random noise and it can also include noise caused by the analysis method. This affects the peak shape and selection of integration boundaries. That is why a robust method is required to handle noise even if the signal-to-noise (s/n) ratio is low. The s/n ratio presents the ratio between the peak level and the noise level in the signal.

The precision of the LC-MS system and peak integration can be inspected with quality control (QC) [5]. This is done by measuring QC samples with the LC-MS system and peak integration method. With the usage of QC samples, it is possible to measure accuracy and precision of the system with metrics. The QC samples have known target values which makes the accuracy measurements possible. The precision of LC-MS system can be measured with relative standard deviation (RSD) of QC sample results. The RSD is calculated from QC measurements performed with the same QC sample in the same laboratory conditions with the same analyzer and method. This provides a metric to compare the precision of different peak integration methods and the performance of individual analyzers.

In order to improve the signal peak detection and integration, different existing algorithms are investigated and a new peak integration algorithm is proposed in this study. The results from the proposed peak integration algorithm are compared to results produced with existing peak integration algorithm and to the current integration method used in the analyzers providing data for this study. In LC-MS peak detection, time, mass, and intensity are measured from the signal [4]. Analyte peaks are usually larger than noise peaks in chromatographic data which makes the separation of analyte peak possible. However, the peaks always include some noise which causes the deviation in the obtained results. A robust peak detection and integration algorithm need to able to adjust the integration limits and area according to the noise level. For example in some of the peak integration algorithms smoothing for the signal is performed prior to the estimation of integration limits.

The peak detection from the signal usually consists of smoothing, baseline correction and peak picking [6]. The open source peak detection algorithms can be separated based on the different methods used in each step. Common smoothing methods used in open source peak detection algorithms are Gaussian filter, moving average filter and Savitzky-Golay filter. Typical baseline estimation functions used in peak detection algorithms are e.g. linear interpolation, moving average minima. The estimated baseline is usually subtracted from the signal. Several criterion are used for obtaining the start and end points of a peak. These include defining a intensity threshold, finding local maxima and selecting N points around it, selecting based on s/n ratio, obtaining ridge lines with continuous wavelet transform and based on peak width. Different methods in these phases can be varied to obtain peak start and end points for the integration.

This thesis has two parts. The first part consists of literature review of the LC-MS method and peak detection methods used in analysis of chromatography data. Section 2 provides introduction to liquid chromatography and mass spectrometry. Individual review is provided for both methods, as well as combination of these two methods. The section 2 also covers review of data processing of LC-MS and discusses quality metrics used in analysis of LC-MS data. In section 3 the used integration methods of open source peak detection algorithms are investigated. The integration method selected for comparison of the proposed method in this thesis is discussed in more detail. The section 3 also discusses theoretical peak fitting to chromatogram data, baseline estimation of chromatogram data and non-linear least square optimization. The proposed peak integration algorithm in this thesis is based on fitting a theoretical peak to the chromatograms with the usage of non-linear least squares.

The second part of the thesis consist of experimenting with LC-MS data obtained from multiple analyzers. The experiments consists of development of pipeline handling raw files obtained from measurements with the LC-MS method, proposing a new peak integration method and experimenting with a known peak integration method from the literature. In section 4, the peak integration pipeline is discussed in more detail. This includes the processing of raw files, reorganizing the processed data, the proposed peak integration method, baseline estimation as part of the proposed method, adding integration method from literature as part of the pipeline and execution of the pipeline. The section 5 includes result analysis of QC runs with multiple QC sets. The results are produced with the peak integration pipeline with multiple integration methods. In the section 6, the conclusion, observations and results in the thesis are concluded.

2 Liquid chromatography - mass spectrometry

Liquid chromatography - mass spectrometry (LC-MS) method combines advantages of liquid chromatography and mass spectrometry. The liquid chromatography method is performed first in the LC-MS method. With liquid chromatography it is possible to separate compounds from a complex mixture [7]. The separation occurs in chromatographic column, where different solvents used in chromatography react with the sample. The different compounds react with the solvents differently, which causes them to flow through the column in different times. In liquid chromatography, the identification of compounds is based on these times, which are called the retention times [7]. However, liquid chromatography is not able to separate components with high confidence when the retention times of components are close to each other.

The number of compounds is so large that it is not possible to be certain that two components are the same with liquid chromatography [7]. With mass spectrometry it is possible to identify compounds with high confidence when the retention time of the compounds in the mixture are close to each other [7]. Compounds with retention time close to each other have different mass spectra which is why mass spectrometry is able to identify compounds with high certainty. However, mass spectrometry is not suitable method for identifying minor components of mixture when the sample includes a lot of different compounds with varying retention characteristics. Mass spectrometry can be used more effectively when the retention characteristics of compounds are close to each other. With combining liquid chromatography and mass spectrometry it is possible to detect compounds with high precision, liquid chromatography separates components of the mixture and retention times are obtained after which mass spectrometry can be used with rather good precision.

However, LC-MS method is not always able to analyze compounds of a sample. That is why LC-MS/MS (tandem mass spectrometry) method have been developed. These tandem mass spectrometers are able to analyze fragment ions of precursor ions which form mass spectral tree of the component [8]. This ability makes the LC-MS/MS method more robust to analyze wider range of samples because different compounds might have the same mass, but their fragment ions have different mass. Sometimes even LC-MS/MS method is not able to fully identify the molecular structures in untargeted metabolomics. This is cause the number of metabolites is so large that is it difficult to find a good match for the gained mass spectra from the literature.

2.1 Liquid chromatography

In liquid chromatography, components of a mixture are separated based on their chemical ability to separate in stationary phase and mobile phase [9]. The affinities of the components for the two different phases cause the components to separate. Some of the components in the mixture are separated by the mobile phases more easily because the mobile phase attracts the components more. On the other hand, the stationary phase retains some of the components which is why they take longer to separate from the mixture. The time that the compound takes to separate from

the mixture is called retention time.

In liquid chromatography the samples need to be introduced to the chromatogram columns where the chromatography occurs. A loop injector is a common tool used to introduce a sample to chromatogram columns [7]. With the use of the loop injector, it is possible to maintain the mobile phase of liquid chromatography and inject samples to the solvent flow of the mobile phase. The flow rate of mobile phase can be set to a certain flow rate and samples can be injected to the solvent flow at certain time intervals.

The components of the sample react with the solvents in the mobile phase and separate from the sample based on their polarity [7]. More polar analytes separate from the sample in the mobile phase because the mobile phase is more polar than the stationary phase. The flow of the mobile solvent can be achieved with gravity or a vacuum [9]. One recent application of vacuum assisted flow is solid-phase extraction where the flow goes through a plastic cartridge with packed particles. The function of the cartridge is to reduce pressure because the LC columns are usually sensitive to handle high pressures.

The liquid chromatography can be normal-phased or reverse-phased [9]. In normal-phased liquid chromatography non-polar analytes elute first from the sample because the mobile phase solvent is non-polar. Normal-phased LC includes polar stationary phase. In reverse phased liquid chromatography, the polarity of phases are the other way around which leads to most polar analytes eluting first from the sample. After the stationary and mobile phases of chromatography the analytes continue to detector system. In LC-MS system analytes continue to the mass spectrometer.

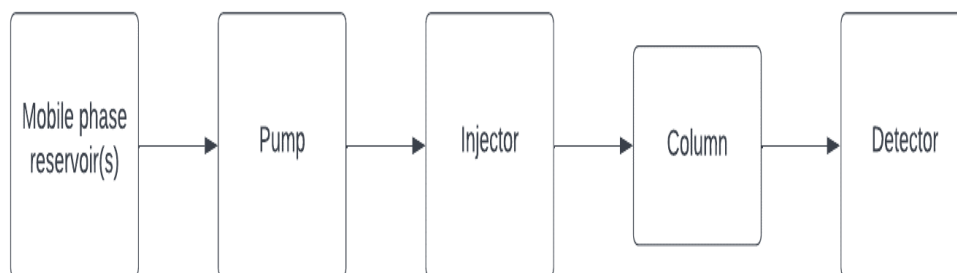


Figure 1: An example of LC system. Edited from Ardrey [7]

2.2 Mass spectrometry

The purpose of mass spectrometry in LC-MS system is to gain structural elucidation of an analyte [9]. A mass spectrometer returns mass-to-charge-ratio (m/z) values of various analytes from ionised analyte molecules. The m/z ratio is calculated by dividing the mass of the molecule by its electric charge. The analytes are ionised so that the m/z ratio and ion abundance can be detected from the ions and fragmented

ions, as MS is not able to measure neutral molecules. With this method it is possible to measure the molecular weight of an analyte. Sample analytes that are received from the LC phase are ionised into an ion source. Introducing the sample to the ion source, ionisation process in the ion source and ion analysis in the mass analyzer are performed in a certain way depending on the technologies that the ion source and mass analyzer use in the mass spectrometer [10]. That is why it is important to know what kind of mass spectrometer is suitable for the planned usage, there are multiple different options for mass spectrometers.

One type of ion source of a mass spectrometer is an electrospray ionisation source (ESI). ESI is a robust ion source that is suitable for polar molecules such as metabolites, xenobiotics and peptides [10]. With the usage of charged capillary between 3 to 5kV of the ESI, a spray of charged droplets are formed at the tip of capillary. Then the droplets are evaporated in order to transfer electrical charge to measured analytes. After the electrical charge is transferred, the analytes move forward in the mass spectrometer through series of small apertures and focusing voltages[11]. ESI is a "soft" ion source which is why it transfers rather small amount of energy to the analyte. That is why ESI does not cause much fragmentation of analytes. In some cases, the fragmented ions are the ions that want to be measured. When ESI method is used, analysis of fragmented ions is possible with MS-MS analysis [12]. The amount of fragmentation can be increased with increasing the voltage of the source. Increasing the amount of fragmentation is suitable for some analytes such as some biological molecules which for the ESI is most widely used [12]. However, this kind of analysis can only be performed in targeted analysis where the m/z ratio of precursor ions are known.

There are also other type of ion sources such as atmospheric pressure chemical ionisation (APCI) and Atmospheric pressure photo-ionisation (APPI) [10]. In APCI the sample travels to the tip of capillary similarly as in ESI. Then the solvent molecules and gas present in the ion source are ionised with a corona discharge. After the solvents are ionised, the ions react with the analytes and ionise them. The APCI is better for small and thermally stable molecules than ESI. In APPI photons are used to ionise molecules. Neutral compounds such as steroids have been analysed with APPI technique.

As mentioned before there are also different kinds of mass spectrometers. One option for mass analyzer is a triple quadrupole tandem mass spectrometer. A triple quadrupole mass spectrometer was used in the process generating data for this thesis. One quadrupole has four parallel metal rods which makes it possible to use varying radio frequency voltages to transmit ions with certain m/z values along the rods [10]. With different voltages it is possible to scan a certain m/z range. The first and third quadrupoles can measure a specific m/z value(s) which makes it possible to improve detection of target analytes because the mass analyzer have more time to detect specific analyte(s) rather than scanning multiple m/z values.

The advantage of the triple quadrupole mass spectrometer is that it offers much more specific mass analysis than single stage mass analysis. With tripe quadrupole mass spectrometer it is possible to analyze precursor and product ions [10]. Product ion is fragmented from the precursor (analyte) ion. This is possible because the

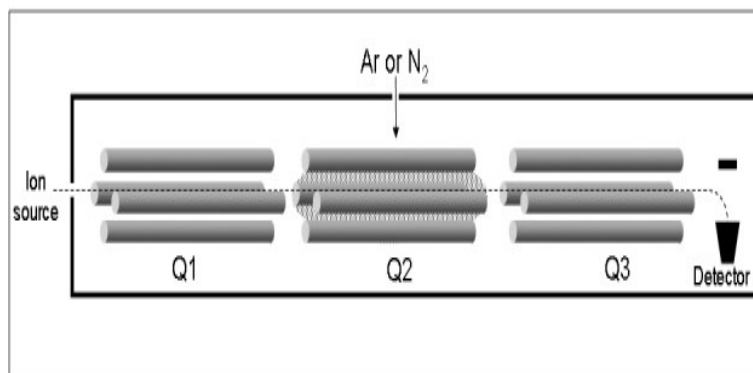


Figure 2: Structure of a triple quadrupole mass spectrometer. Edited from Pitt [10]

second quadrupole functions as a collision cell where the product ions are fragmented with collision gas such as argon or nitrogen. The m/z values of precursor ions are scanned in the first quadrupole after which the precursor ions travel to the second quadrupole where they are fragmented. After fragmentation the product ions move to the third quadrupole where their m/z values are scanned. This kind of functionality is good for analytes that have a common m/z value for precursor ion but more rare m/z value for product ion [10]. For example 25-hydroxy vitamin D3 precursor ion has mass of 401 m/z which can be quite common in biological sample. One of the fragmented product ions of 25-hydroxy vitamin D3 have the mass ratio of 383 m/z . Knowing that the probability of having some other ion among the fragmented ions with the same exact 383 m/z ratio is low, makes it possible to analyse 25-hydroxy vitamin D3 with the LC-MS/MS method.

In tandem mass spectrometry, the chromatogram is formed by the mass transition between precursor and product ion. Transition of multiple ions are measured in the analysis process. When measuring a specific analyte, the measurement can be referred as quantifier transition [13]. This transition forms the chromatographic peak used in the result analysis. Quantifier transitions are often qualified with qualifier ion transitions in case of interference. Quantifier and qualifier ions have the same precursor ion but different product ion, which makes the qualification possible [13]. The concept of qualifier transitions are used in the analysis process in this thesis and the result analyte transition is referred as quantifier transition.

There are also other types of mass spectrometers such as the time-of-flight (TOF) analyzer and ion trap analyzers [10]. In TOF analyzers ions are guided through the mass analyzer with high voltages. Because of the high voltages, ions travel through the TOF analyzer in different times because of their different electrical charges. That is how the m/z values are measured with TOF analyzer. The functionality of ion trap analyzers is based on ejecting ions from the ion trap based on their m/z value in order to form mass spectrum [10]. After the ions are analyzed with a mass analyzer, the ions travel to the detector which is the last part of a mass spectrometer. The detector is used to measure abundance of the analytes.

2.3 Data processing

The data obtained from LC-MS process is complex which is why it needs processing before it can be analyzed [4]. The LC-MS data consist of raw files which contain the MS spectra. This raw data has three dimensions which are signal intensity, retention time and mass-to-charge ratio. One raw file usually include the signal intensity as function of retention time. The data set consist of similar raw files where each raw file is from a specific m/z value, forming a range of raw files over a measured m/z values. The raw files can be processed by processing features individually and not taking other features into account or by using divided bins of the signal and forming recognition profile for each sample. The selected approach usually depends on the research question.

The data processing usually has three phases [3, 4]. In the first phase, the noise of the signal is reduced with a filtering method. This signal filtering can be done with different smoothing methods such as moving median or mean filters. The filter should not eliminate parts of the relevant peak in the chromatogram because it could alter the result. The noise reduction phase is optional but reducing noise from the signal can lead to better results if the noise is reduced successfully without altering the signal peak.

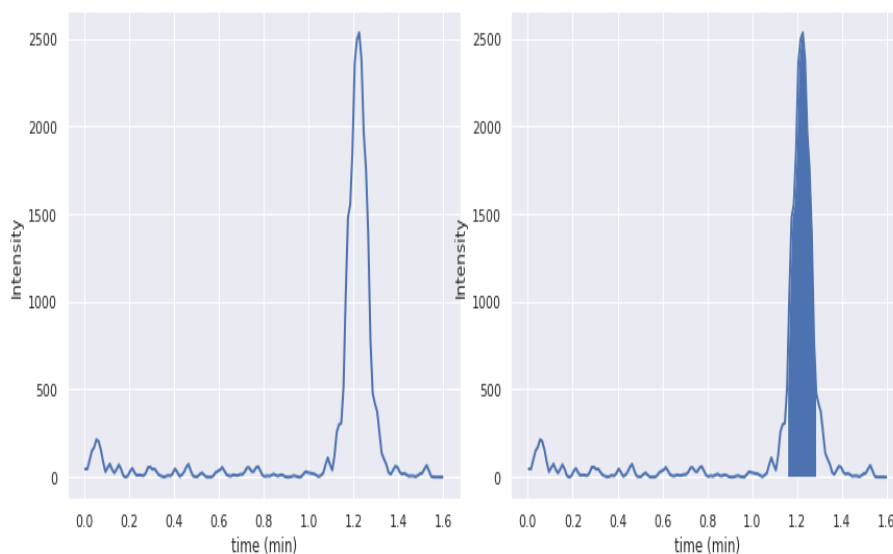


Figure 3: An example of chromatogram signal and integrated chromatogram signal. The highest peak is the peak from the analyte and other peaks are noise in the signal. The goal of the peak detection algorithm is to detect the peak from the signal and remove noise from it in order to get the precise peak area.

Peak detection from the signal is done after noise reducing. Multiple peak detection algorithms have been developed [4, 14, 15, 16, 17]. Peak detection algorithms have two main aspects that need to be considered. The algorithm needs to be able to differentiate analyte peak from noise peaks and extract characteristics of the analyte peak. There are different approaches how the peak detection algorithm can differenti-

ate analyte peak from the signal [18]. These include for example analyzing the shape of peak, the length of peak or calculating the peak intensity. These approaches can also be combined in a peak detection algorithm. The used peak detection algorithm provides points for integration of the detected peak. The peaks are integrated and the calculated areas represent the result of each measurement.

After peak detection the peak alignment can be performed. [3, 4]. Corresponding peaks of the same analyte are matched and compared in peak alignment phase. In some cases, only the integrated areas of the peaks are compared. There are different peak alignment algorithms for comparing the peaks to each other. Peaks that are from the same analyte have close m/z values in signal but retention times can vary between runs [3]. That is where peak alignment is needed to make accurate comparisons. Peak alignment faces multiple problems [19]. For example retention times of the same analyte can differ between runs, peaks may differ because of drifts, and single analyte could be detected as multiple peaks or the analyte may be missing from the sample. After peak alignment is done, the peaks should be normalized in order to reduce systematic error and improve the data quality [3, 19]. After these steps are done data can be visualized and analyzed statistically.

2.4 Quality metrics

Even after all the data processing steps of LC-MS data have been performed the data includes low quality peaks. These low quality peaks can have low s/n ratio or missing values which makes the quality and reliability of the data worse. Quality control (QC) samples can be used to assess the quality of data [5]. With QC samples it is possible to describe quality of gained data by converting analytical accuracy, precision, and repeatability of QC samples to metrics. With these metrics it is possible to flag or remove low quality features.

In this thesis we are interested in the measurement error that happens in LC-MS analysis. The measurement error can be caused by random error or systematic error [5]. The systematic error is caused by impaired analytical method or instrument but it can be controlled with an internal standard. Internal standard samples include a known concentration of an analyte which is why they can be used to quantify the analytes measured in a sample. Internal standard are used in this study to reduce the systematic error and they include high concentration of measured analytes which leads to high signal-to-noise ratios. However, random errors can not be avoided. Random errors can be cause variation of factors between analyses without knowledge of the reason. Thus it can be reduced by optimizing the used analytical methods and instruments.

The precision of the measurements can be measured by running multiple QC samples and comparing the individual results. The QC samples need to be from the same mixture and the runs need to be performed with the same method, instrument and equipment in the same laboratory [5]. The precision of the runs can then be calculated with standard deviation of the results of the QC runs. In order to compare the precision between different analytes and instruments, relative standard deviation

(RSD) is used. RSD can be calculated from

$$RSD = \frac{s}{m} * 100\%, \quad (1)$$

where s is standard deviation and m is mean of one set of QC runs [5]. There are also other metrics to measure the precision of QC runs. Median absolute deviation (MAD) is commonly used in untargeted metabolomics to derive robust estimate for RSD [5]. Scaling factor of 1.4826 is used in MAD [20]. Mad can be calculated from

$$RSD^* = \frac{1.4826 * MAD}{median(x)} * 100\%,$$

where x includes the measured peak areas of QC samples and MAD is the calculated median absolute deviation. MAD is non-parametric statistical equivalent to standard deviation. Alternative metric for measuring precision of QC runs is D-ratio. D-ratio measures for example variability or spread between pooled QC samples and biological test samples [5]. The D-ratio can be calculated with standard deviation or MAD of test sets from formulas

$$D-ratio = \frac{s_{qc}}{s_{bs}} * 100\%,$$

where s_{qc} is standard deviation of QC samples and s_{bs} is standard deviation of biological test samples and

$$D-ratio^* = \frac{MAD_{qc}}{MAD_{bs}} * 100\%,$$

where MAD_{qc} is median absolute deviation for QC samples and MAD_{bs} is median absolute deviation for biological test samples. Small result from all the formulas above indicate better precision of the measurement method. In this thesis RSD is used to measure the precision of the measurement sets. The result of an individual sample run is measured by calculating the peak area that each individual run forms. The RSD of these areas is then calculated. With these metrics it is possible to measure the quality of set of runs with different analytes. If the gained metric is over certain threshold the run set can be removed from the whole data set which improves the quality of the data. Usual threshold for accepting set of measurement with RSD is $< 20\%$ [5]. In this thesis, the goal is to compare RSDs of different results obtained with different peak detection and integration procedures.

3 Peak detection and fitting algorithms

The aim of peak detection in LC-MS is to detect the peak from the chromatogram signal produced by the LC-MS method. The signal peak need to be integrated from the chromatogram signal in order to get the results from the method. Determining the peak limits for the integration of the signal peak is hard because of the noise included in the chromatogram signal. The noise in the signal causes variation in the peak shape and amplitude which makes the estimation of the integration points difficult. This section introduces methods for detection and integration of peaks. These include peak detection algorithms finding the peak from the signal, determining points for integration and integration of fitted theoretical peaks.

3.1 Open-source peak detection algorithms

Multiple open-source methods have been developed for LC-MS data processing and peak detection such as XCMS [4], OpenMS [14], MZmine [15], MS-DIAL [16] and MAVEN [17]. These packages offer tools for processing and investigation of LC-MS data such as peak visualization, identification of peaks, retention time alignment and raw file processing. However, in this thesis we are only interested in the peak detection and integration methods of these open source packages.

In the XCMS package, the LC-MS data is divided to extracted ion base-peak chromatograms [4]. These chromatograms include an individual peak which is detected and integrated from the chromatogram signal. In this study we experiment with similar data, each chromatogram includes only one peak. The peak detection and integration method in the XCMS package is based on filtering the signal with the second derivative of Gaussian to obtain boundaries for the integration. Obtaining the boundaries with the second derivative of Gaussian is explained in more detail in the next chapter.

The peak position in the OpenMS [14] package is defined with continuous wavelet transform. The continuous wavelet transform method is explained in more detail in the next chapter. The used wavelet function in the continuous wavelet transform method is Marr-wavelet function [21] instead of the commonly used second derivative of Gaussian. The Marr-wavelet takes the form

$$\psi_x = (1 - x^2) \exp\left(\frac{x^2}{2}\right), \quad (2)$$

The maximum point of the wavelet signal is considered as the maximum position of the chromatogram peak. After wavelet transform, the peak end points are estimated. This is done by moving left and right from the peak maximum. When minimum is found or the intensity is below threshold for the noise defined prior peak detection, the peak endpoint is found [21]. After finding the peak end points, several peak parameters are estimated with fitting asymmetric Lorentzian and hyperbolic secant functions to the chromatogram peaks. These parameters include height, peak area, full-width-at-half-maximum and signal-to-noise ratio. The asymmetric Lorentzian

takes the forms of

$$L(x) = \frac{h}{1 + \lambda^2(x)(x - c)^2}$$

and hyperbolic secant takes the form of

$$G(x) = \frac{h}{\cosh(\lambda(x)(x - c))^2},$$

where $\lambda(x) = \lambda_l$ if $x \leq c$ and $\lambda(x) = \lambda_r$ if $x > c$. λ_l and λ_r are the left and right widths of the peak [21]. h describes the height of the peaks in the functions. The peak area is obtained by integrating from left end point to the peak to the centroid and from centroid to the right end point of the fitted peak. Nonlinear optimization can be used to optimize the parameters obtained from the fitting in the OpenMS package.

The MZmine [15] package connects chromatograms of the same m/z value together from multiple measurements of multiple different analytes and detects peaks from them. After connection, each chromatogram is deconvoluted into individual peaks. Each individual peak is detected from the chromatogram with a baseline cut-off and noise amplitude algorithms. If a peak intensity is over the baseline cut-off and noise amplitude thresholds the peak is detected from the chromatogram. After the peaks are detected, the Savitzky-Golay algorithm is used to detect the boundaries of each peak. Smoothed second derivative of the chromatogram curve is used by the algorithm to detect the boundaries of the peak.

The MS-dial [16] package also performs smoothing of the chromatogram signal. This is done with linearly weighted smoothing average method. The method can be expressed with form

$$f(x) = \frac{\sum_{i=-m}^n (n - i + 1) \times f(x + i)}{n^2},$$

where n is the number of data points [16]. The MS-dial package offers also optional smoothing algorithms such as moving average, Savitzky-Golay and binomial filters. After smoothing, three thresholds are determined for the peak detection. These include the maximum differences between two adjacent points, the maximum of first derivatives and maximas of second derivatives in the chromatogram. Five-point approximations are used to obtain the derivatives. These are calculated from values which are 5% below each maximum. Then the left endpoint of the peak is obtained when the amplitude of the first derivative exceed amplitude threshold and first order derivative threshold in two adjacent points. The right endpoint of the peak is then found with a similar method.

Peak detection in the MAVEN [17] package is performed by dividing the data to extracted ion chromatograms (EICs) with binary search based on m/z value and retention time. One EIC include one chromatogram peak from specific ion. The baseline for the peak detection is calculated with Gaussian smoothing. 20% of the highest intensity values of the EIC is removed prior the smoothing. Gaussian filter

takes form

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (3)$$

The median of the smoothing results is then considered as the baseline. The peak boundaries are then detected by the sign of the derivative. If the sign of the derivative changes, the boundary point is detected. The boundary point is also detected if the intensity is below the defined baseline.

3.2 Continuous wavelet transform

Continuous wavelet transform (CWT) can be used as part of a peak detection algorithm. The CWT algorithm transforms the signal into wavelet space. The wavelet coefficients from the wavelet space provide a way to present shape information of peaks [22]. Normalised second derivative of Gaussian is used as the wavelet in the method. Multiple peak detection methods such as XCMS [4], peak detection algorithm developed by Du et al. [22] and centWave [23] take advantage of second derivative of Gaussian in the peak detection. One advantage of using the wavelet transform is that no peak smoothing or baseline removal or is needed for the signal. Peak detection algorithms usually require baseline removal and smoothing which can lead to removal of parts from the real peaks in the signal [22]. The CWT method is a robust method because it doesn't require smoothing or baseline removal from the chromatograms.

The CWT method is suitable for detecting peaks with different width and height. This is possible by performing the CWT method with multiple scales. Analytes with high m/z tend to have lower amplitude and be wider [22]. That is why it is important that the method can detect peaks with different amplitude and width. The mathematical formula of the CWT algorithm is:

$$T(s, \tau) = \int_{-\infty}^{\infty} f(t)\psi_{s,\tau}(t)dt, \psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right), s \in \mathbb{R}^+ - 0, \tau \in \mathbb{R}, \quad (4)$$

where τ is the translation, s is the scale, $f(t)$ is the signal, and ψ is the mother wavelet [23]. Result T from the algorithm is a two dimensional matrix which includes wavelet coefficients. Normalized second derivative of Gaussian probability function is usually used as a mother wavelet in the algorithm [22, 23].

The analyte peak is compared to the scaled mother wavelet. If the scale is $s = 1$ the algorithm provides optimal matches for analyte peaks that have width of two sample intervals. If the scale is increased to s_2 the CWT method provides best fit for peaks with width $2s_2$ [22]. The mother wavelet, also called Mexican Hat wavelet, can be seen in figure x with multiple different scales. From the 4 it can be observed that the local maxima of CWT is around the centre of the signal. The start and end point of the signal can be obtained from the zero crossing points of the Mexican Hat wavelet. The same method is applied in OpenMS [14] but the wavelet function used is Marr-wavelet 2.

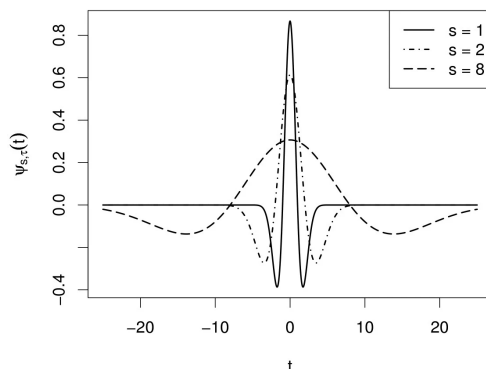


Figure 4: Mexican Hat wavelet with different scales. Edited from Tautenhahn et al. [23]

Including the CWT method as part of peak detection can be done with in different ways described in more detail here. In the beginning of the CWT method the chromatogram signal is processed with the CWT method with multiple scales using the formula 4. This provides CWT coefficients from the signal $f(t)$ using different scales. The CWT coefficients are in the wavelet coefficient matrix T formed by the CWT method Tautenhahn et al. [23]. The matrix has $M \times N$ dimension, N is the number of scales used and M is the length of the data. Each row in the matrix is the signal derived with Mexican Hat wavelet which is comparable to the second derivative of the Gaussian with one of the scales. In the XCMS [4] the best results in the study were obtained when the width of the derived model was 1.5 – 4 times the original signal peak width.

In the peak detection algorithm developed by Du et al. [22] and centWave [23] local maximas of CWT coefficients at each scale level are identified. This can also be called ridge detection. Ridge is a scale range in which the chromatographic peak is located. The ridge detection is done with the local maximas which are linked by forming ridge lines. This is done by looping through the CWT coefficient matrix starting from the row with the highest scale. The steps of forming the ridge lines between local maximas is explained in detail in Du et al. [22]. From these ridge lines it is possible to get the scale range where the chromatographic peak is located. Three rules are applied to the ridge lines to obtain the peaks from the chromatogram [22]. The maximum amplitude on the ridge line with the certain scale should be in a defined range, signal to noise ratio should cross a certain threshold and the length of the ridge lines should cross a defined threshold [22]. From these conditions the peak position can be estimated in two ways. Ridge line can be followed from high scale to small scale. Eventually some scale point includes the estimation for the position of the peak maximum. Analysing maximum CWT coefficients on the ridge line at specific scale range can be used as other way is to estimate the peak position. In Tautenhahn et al. [23] study, start and end points of the chromatogram peak are estimated with the CWT coefficients of the scale which provides most optimal localization of the peak.

3.3 Fitting a theoretical peak

In Stevenson et al. [24] study, integration of chromatogram peaks was performed by fitting theoretical peak functions to chromatograms. The relative standard deviation of integrated (RSD) areas was compared to RSDs of integrated areas obtained with manual determination of integration boundaries, examining the relative signal intensity compared to the noise of signal to obtain integration boundaries, analysis of signal slope to get the boundaries and using the signal width as a standard function to obtain the boundaries.

In Stevenson et al. [24] study, 4-parameter Exponentially Modified Gaussian (EMG) and EMG + half Gaussian modified Gaussian (GMG) models were fitted to the experimental data. The mathematical expression of these functions can be found from [24, 25]. In addition to EMG and GMG models, basic Gaussian 5 and parabolic-Lorentzian function (PLMG) 6 were fitted to the data. PLMG is given by

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}} \quad (5)$$

$$f(x) = H_0 \exp\left(-\frac{1}{2} \frac{(x - x_R)^2}{\sigma^2}\right) \quad (6)$$

where

$$\sigma^2 = \sigma_0^2 + m \frac{(x - x_R + d)^2}{1 + \left(\frac{(x - x_R + r)^2}{w^2}\right)}. \quad (7)$$

Here x is the time, $f(x)$ is the detector response, H_0 is the maximum of the peak at x_R and σ is the peak width [24]. σ_0 , m , and d are parameters from parabolic function and r and w are parameters of the Lorentzian function. In Gaussian function 5 a is the amplitude of the peak, b is the location of the peak centre and c is the standard deviation.

The models above were fitted into experimental data which consisted of 10 replicate measurements of one analyte forming 10 individual peaks to fit the models to [24]. The integration of the models were performed from zero to infinity. The quality of the fit was calculated with formula

$$\epsilon_r = \frac{\sum_{i=1}^N |h_i - \hat{h}_i|}{\sum_{i=1}^N |h_i|} \times 100.$$

The PLMG model provided the best fit for the data with ϵ_r of 2.34%. Gaussian model, GEMG4 and EMG + GMG models provided fits ϵ_r of 7.53%, 7.70% and 4.99%. Because the PLMG model provided the best fit for the data, it was selected as part of further comparison to different integration methods in the study. The RSD of the integrated areas with PLMG method was 0.44%. In comparison to other methods, the RSD calculated with the PLMG method outperformed manual determination of integration points which was 0.51%. However better RSDs were obtained with the gradient analysis method, the peak width analysis method and the relative signal intensity method which were 0.39%, 0.43% and 0.32%.

Fitting a theoretical peak to experimental data have been used to determine integration boundaries for the chromatogram peaks. This is done in Stevenson et al. [26] study by fitting EMG algorithm to experimental data to obtain parameters for further determination of the peak boundaries. The EMG algorithm is given by:

$$f(x) = a \exp\left(-\frac{(x-b)^2}{2c^2}\right) \times \frac{1}{\tau} \exp(-x/\tau). \quad (8)$$

The EMG function is combination of Gaussian peak (5) and exponential decay function where τ is an exponential modifier. The exponential modifier is used to measure difference between ideal symmetrical peak and actual peak shape. The EMG provides parameters for peak width multiplier method determining the integration boundaries in the study.

The peak width multiplier method is based on peak width at half height $w_{1/2}$. To get the whole width of the peak, the $w_{1/2}$ is multiplied with width multiplication factor n [26]. This can be described with equation:

$$w_{peak} = w_{1/2} \times n. \quad (9)$$

The width multiplication factor is determined by analyst and w_{peak} is the boundary peak width. However, real peaks are rarely symmetrical which is why peak start and end times are defined by multiplying 9 with an arbitrary weighting function given by:

$$t_{start} = t_R - w \times w_{peak} \quad (10)$$

and

$$t_{start} = t_R + (1 - w) \times w_{peak}. \quad (11)$$

The t_R in the functions 10 and 11 is the peak centre.

Accuracy of integration boundaries obtained for real peaks was not reported in Stevenson et al. [26] because first absolute and the second central moments were not available for real peaks. However, visually acceptable boundaries were obtained. The peak width multiplier method was further tested on simulated peaks. The simulation was performed by adding baseline noise to the theoretical peaks. Gaussian and EMG model was used to create the simulated peaks, and the real second central moments of the peaks were compared to the second central moments of the peaks with added noise. The moments were then obtained by varying the peak width multiplier and weighting factors. Different amount of baseline noise was also applied in the experiments.

The results from the study [26] show that the peak width multiplier method obtained accurate measurements for integration boundaries when the signal to noise ratio was over 200 for the Gaussian peak with added noise. The RSD for obtaining the moments were lower when the signal to noise ratio was higher. With the EMG peak a linear relationship between the width multiplier, the degree of peak tailing, and the weighting function was found. However, optimum for the integration parameters was no found. The peak width method was also tested in Stevenson et al. [24] study

as mentioned before in this chapter. Their width multiplication factor n in (9) was defined by an analyst in each peak integration. The RSD of the integrated areas was rather good compared to other integration methods in the study.

Peak fitting to chromatogram data was also performed in Nikitas et al. [27] study. In the study, several peak algorithms were fitted to the data with non-linear least squares fitting. The position of the peaks and the height of the peak was calculated for the fitting in order to make the fitting function better. Multiple different peak functions were fitted, and the goodness of fit was estimated with sum of squares of residuals. The fitted functions included asymmetric Gaussian distribution function (AGD) 5 where $c = c_1$ when $x < b$ and $c = c_2 \neq c_1$ when $x > b$, Generalized exponential function (GEX) given by

$$h(t) = \begin{cases} h_m \left(\frac{t-t_0}{t_m-t_0} \right)^{b-1} \exp\left(\frac{b-1}{a} \left(1 - \left(\frac{t-t_0}{t_m-t_0}\right)^a\right)\right) & t > t_0 \\ 0 & t \leq t_0, \end{cases}$$

where a and b are constants, t_0 is starting time of the chromatogram signal, t_m is the position of peak maximum and h_m is the height of the peak, Lorentzian function (L) given by

$$h(t) = \frac{h_m}{1 + (t - t_m)^2/s^2} \quad (12)$$

where s is the standard deviation, polynomial modified Gaussian (PMG) where standard deviation has the form

$$s = s_0 + s_1(t - t_m) + s_2(t - t_m)^2 + \dots, \quad (13)$$

EMG 8 and generalized EMG (GEMG). The difference between (7) and (12) is that (12) is always symmetric, and (7) can be asymmetric curve.

Peak fitting was performed with all these functions in the study [27]. For AGD, GEX and L the sum of squares of residuals indicated poor fits for the chromatograms and they were not investigated further. The other functions were fitted to multiple data sets. The PMG and GEMG functions were able to describe all chromatographic peaks well, and EMG was able to describe a part of the chromatograms well. Two version of PMG functions were fitted which both provided good fits. In the first version of the 13 $h(t \rightarrow \inf) = h_m$ and in the second version $h(t \rightarrow \inf) = 0$ This causes differences at the end of the functions. It was concluded that if the functions is one of the ones with good fits above, the choice of the function does not play that important role, all the functions can describe a chromatographic peak rather well.

3.4 Baseline estimation

Chromatogram measurements include background and noise in addition to the chromatogram peak. Here the term baseline refers to the noise and background part of the chromatogram. The goal of baseline approximation is to approximate the baseline as well as possible, in order to take it into account in peak detection and

integration. Baseline estimation is also related to smoothing of the chromatogram, which is done prior to the peak detection in many peak detection algorithms. It is up to the developers to decide if baseline estimation has the potential to improve their peak detection and integration.

The noise in the chromatogram can significantly affect the functionality of algorithms detecting and integrating the chromatogram peak [28]. This problem is not just restricted to chromatogram signals, it is also related to other signal processing. Multiple baseline estimation methods have been developed tracing all the way back to the late 1970s. Pearson [29] developed baseline estimation algorithm based on separating peak points and baseline points. This is done by comparing the intensity of a point to standard deviation interval. This method requires definition of parameters for the interval and poor selection of these parameters could lead to errors.

Moore and Jorgenson [30] developed a simple baseline estimation method which is based on median filtering. Moving median filter was used to remove noise spikes from the chromatogram signal. This method can successfully be applied to chromatograms with narrow peaks with wide baselines. With broad peaks the method estimated the baseline badly. Another filters have also been applied into signals to estimate baseline such as Kalman filter and wavelet-based filtering [28]. Applying wavelet based filter can be used also to estimate peak position as described in chapter 3.2.

Iterative methods have also been used to estimate the baseline of a signal. Ruckstuhl et al. [31] introduces robust baseline estimation technique which removes baseline from spectral data. The method defines the spectrum sum of baseline signal, baseline-free signal and measurement error. These three components are separated in order to get the baseline. Also, polynomial smoothing [32] and reweighted Whittaker smoothing [33] have been performed with iterative methods. [28] also proposes an iterative baseline estimation method based on skewness of the residuals.

In this study, we experiment with one-dimensional Sensitive Nonlinear Iterative Peak clipping (SNIP) algorithm for baseline estimation. The SNIP algorithm takes form of

$$v_p(i) = \min[v_{p-1}, 0.5(v_{p-1}(i+p) + v_{p-1}(i-p))], \quad (14)$$

where $v(i)$ is the signal at point i and p is the iteration [34]. The baseline value at point i is obtained step by step from formula 14 by comparing the average of previous and next values in the signal and the previous value in the signal and taking the minimum value of those. Before applying the signal $v(i)$ to the formula 14, LLS operator is applied to the signal. The LLS operator takes form of

$$v(i) = \log[\log(\sqrt{y(i)} + 1) + 1],$$

where $y(i)$ is the original signal. The square root in the formula enhances small peaks and the log makes it possible to work over a few orders of magnitude. The SNIP algorithm is considered to be an efficient method for baseline detection because it can determine peak-free regions and peak regions.

Another more recent baseline and noise estimation algorithm BEADS (baseline estimation and denoising using sparsity) [35] was also taken into experiments in

addition to the SNIP algorithm. Signal can be considered as sparse if it can be described with a few non zero parameters, which is the case when the chromatogram includes a few peaks. In Beads, the chromatogram signal is modelled as

$$y = x + f + w,$$

where x includes the peaks, f includes the baseline and w the noise. The baseline is estimated with a low-pass filter

$$f = L(y - \hat{x}),$$

where \hat{x} is an estimate of the peak. Then the noiseless chromatogram can be estimated by

$$s = L(y) + H(\hat{x}),$$

H is a high pass filter. The suitable filters are obtained by minimizing two complex cost functions. The Beads algorithm uses the sparse signal and its first two derivatives to estimate the baseline and noise.

3.5 Nonlinear least-squares fitting

Nonlinear least-squares fitting can be used to fit data points to theoretical model. This is done by minimizing sum of squares between set of data points and outputs of a model function. If there is a set of m data points (t_i, y_i) and model function $\hat{y}(t; p)$ with independent variable t and vector p with n parameters, it can be expressed as

$$\chi^2(p) = \sum_{i=1}^m \left(\frac{y(t_i) - \hat{y}(t; p)}{\sigma_{y_i}} \right)^2,$$

where σ_{y_i} is the measurement error[36]. This can be further expressed as

$$\begin{aligned} \chi^2(p) &= (y - \hat{y}(p))^T W (y - \hat{y}(p)) \\ &= y^T W y - 2y^T W \hat{y} + \hat{y}^T W \hat{y} \end{aligned}$$

where W is the weighting matrix [36]. In curve fitting the weights $W_{ii} = \frac{1}{(\sigma_{y_i})^2}$ in matrix W is used as goodness-of-fit measurement for the curve. For nonlinear functions minimization of the χ^2 need to be done iteratively with the model parameters p . At each iteration the goal is to reduce χ^2 by finding perturbation for parameters p .

Minimization of χ^2 and updating parameters of the model can be performed with gradient descent optimization. In this method gradient of the least squares model (6) is formed with respect to model parameters p . This can be expressed as,

$$\frac{\partial}{\partial p} \chi^2(p) = 2(y - \hat{y}(p))^T W \frac{\partial}{\partial p} (y - \hat{y}(p))$$

$$= 2(y - \hat{y}(p))^T W J,$$

where J is Jacobian matrix $\frac{\partial \hat{y}}{\partial p}$ [36]. The sensitivity function \hat{y} which is dependant to variation in the parameters p is represented with the Jacobian matrix. The perturbation h is found by updating model parameters p in "downhill" direction on our error function towards the best fit. In each iteration parameters are updated and they yield slightly lower error than previous iteration. The perturbation h is presented as

$$h = \alpha J^T W (y - \hat{y}),$$

where α determines length of the step in gradient decent direction [36].

The minimization of the nonlinear least-squares method can also be done with the Gauss-Newton method. This method assumes that the object function in least-squared method is quadratic when the parameters of the function are near to the optimal solution [36]. In this method first order Taylor series expansion is used to approximate perturbed model parameters.

$$\hat{y}(p + h) \approx \hat{y}(p) + \frac{\partial \hat{y}}{\partial p} h = \hat{y} + Jh,$$

where h is the perturbation. This equation is substituted to equation (8) which gives us

$$\chi(p + h) \approx y^T W y + \hat{y}^T W \hat{y} - 2y^T W \hat{y} - 2(y - \hat{y})^T W J h + h^T J^T W J h,$$

which leads to quadratic approximation in the perturbation h [36]. h that minimizes χ^2 found by taking partial derivative of χ^2 respect to h and checking when this derivative is equal to zero. By finding the h that minimizes χ^2 we can find the optimal parameters for fitting the model [36]. The partial derivative gives us

$$\frac{\partial}{\partial h} \chi(p + h) \approx -2(y - \hat{y})^T W J + 2h^T J^T W J$$

and the normal equations of the Gauss-Newton are

$$[J^T W J] h = J^T W (y - \hat{y}).$$

It can be observed that the right sides of equations (11) and (15) are similar.

In the experiments in this thesis, the Levenberg-Marquardt method is used to get the optimal parameters from nonlinear least-squares fitting. This method combines gradient descent and Gauss-Newton optimization methods by varying the methods in the iterations [36]. This varying is done with damping parameter λ .

$$[J^T W J + \lambda I] h = J^T W (y - \hat{y})$$

The damping parameter is added as part of Gauss-Newton method. When damping parameter is large the Levenberg-Marquardt method uses gradient decent method. As the gradient decent method proceeds and the solution improves and the damping

parameter is reduced as the result improves. This eventually leads to the method using the Gauss-Newton method. In the case the parameters in optimization are near the optimal values and the gradient decent is able to get the parameters close to the optimal values, the Gauss-Newton method performs well finding the local minima [36]. That is why the Levenberg-Marquardt method starts with the gradient descent method and ends with the Gauss-Newton method.

4 Framework

A peak detection and integration framework was developed as part of this thesis and it was applied to chromatogram data gained from clinical LC-MS analyzers. This section presents the proposed peak detection and integration framework. The motivation behind the peak detection process is explained and the methods for peak detection are presented.

4.1 Motivation

The motivation behind this study is to improve the performance of the peak integration algorithm by reducing the relative standard deviation (Eq. 1) of the obtained peak areas. The analyte peak results are produced with a clinical analyzer using the LC-MS measurement technique. The currently used method for peak detection and integration in the analyzers produces results with moderate RSD, but there were reasons to believe that the current method could be improved. The results obtained with the currently used method are compared to results obtained with peak fitting and integration method developed as part of this thesis, and to results obtained with known existing methods.

The RSD is calculated from responses of 20 peak area measurements which have the same target value. Therefore, the response RSD is an estimate of the imprecision of the instrument. The chromatogram peaks are formed by mass transitions between precursor and product ions. The concept of precursor and product ions is discussed in more detail in chapter 2.2. The peak areas used to calculate the response are from the quantifier and internal standard transitions. In quantifier transition, a selected molecule is measured from the sample with the mass spectrometer. Measuring the internal standard transition works similarly, the concept of using internal standard is discussed in chapter 2.4. Peaks formed by qualifier transitions are also collected. Qualifier transitions are used to qualify the quantifier transition signal, the molecule measured in qualifier transition have the same precursor ion as the measured molecule in quantifier transition, but the formed product ions in the transition are different. That is why qualifier transition can be used to perform quality checks for quantifier transition in case of interference caused by other molecules of the sample entering the mass spectrometer.

One response is calculated dividing the integrated peak area of quantifier chromatogram with the integrated peak area of internal standard chromatogram. The mass spectrometer is used to measure multiple transition of molecules in the analytes, from which the quantifier and internal standard transitions are used to calculate the response (result) of the measurement. The RSDs are measured from a sets of 20 measurement runs producing 20 responses, which are performed with the same quality control sample, i.e. the results with the analyzing method should be the same. However, this is not the case, the results always include some random noise. By reducing the effect that the random noise causes between the measurements, it is possible to produce results with better precision. Better precision in the analysis process would improve the quality of the results produced with the analyzer.

4.2 Software

Python programming language was used to create the pipeline for obtaining the analyte peaks from raw data, matching the measurement sets to the chromatograms, fitting models to the chromatograms, integration of the model peaks and calculating RSDs for the sets of fitted peaks. The language was chosen because it has great data mining, peak fitting and integration tools. Multiple different Python packages were used in this study. These Python packages included NumPy, pandas, SciPy, Matplotlib and seaborn. NumPy package offers comprehensive mathematical functions and fast and versatile vectorization and indexing of arrays. Pandas package offers useful source data manipulation and analysis tools. Scipy package offers algorithms for optimization, integration, statistics and algebraic equations. The Matplotlib and seaborn package offers python visualization tools. The analysis of gained results were performed with Python, Excel and Minitab.

4.3 Data acquisition

The data for this thesis were collected from quality control (QC) runs performed with 30 different analyzers and two different analytes. The used analyzers have two different channels for measurements, both channels were used for measurements. The QC runs were performed in sets of 20 measurements, each of them being analyzer, analyte and channel specific. Each measurement was measured from the same QC sample, which is why the results should be the same in a measurement set. These measurements produce raw files which include the chromatogram signals. These chromatogram signals needed to be acquired from the raw files in order to apply peak detection and integration methods from the literature and the proposed methods in this thesis. The results of the same QC runs produced with the currently used peak detection and integration method were stored in excel files which required separate data acquisition. The excel files were formed by processing the same raw files with Thermo Scientific™ Xcalibur™ software using the current method. The phases of data acquisition from these files is described in this chapter.

4.3.1 File processing

Two different kinds of files needed to be processed to obtain all the needed data. In order to compare the proposed and existing methods to the currently used method, data needed to be collected from the previous RSD studies. This data was stored in excel files including data from the processed raw files of the QC measurements. Each excel file included measurements from one analyzer. The data in these excel files was divided by analyte and channel forming 4 different measurement sets, each set including 20 responses. The responses were calculated from measurements including chromatograms of measured transitions.

One of the measurements included chromatograms of each measured transitions including quantifier, qualifier and internal standard transition. The integrated areas of each signal peak were stored in the excel files from which the responses were calculated by dividing the integrated area of quantifier signal by integrated area of

internal standard signal. From these responses, the RSD of each measurement set was calculated and collected in the data acquisition. The RSDs were collected to a separate array for future comparison. In addition to the RSDs, also other parameters were collected from the excel files in order to form similar quality control sets from the data of raw files as in these excel files. These parameters included JobID, channel of the measurement, analyzer and the analyte name. JobID is a unique identifier for a measurement. One JobID includes chromatogram signals of different ions of the two measured analyte. All these values were collected to a separate array. The data collection from the excel files was performed with Python, with the usage of pandas package.

A Python script was created for processing the raw files. Separate processing of raw files was required because processing the files with Xcalibur™ software did not save the intensity values and measurement times of the chromatograms to the excel files. One raw file included one chromatogram signal formed by a transition of the measured molecule of the analyte. From each processed raw file the Python script collected the intensities forming the chromatogram signal, the measurement times, JobIDs identifying each measurement, compound (transition) names and analyte names. These values were collected from all the raw files and stored to a separate array. In total 7200 raw files were processed to obtain the chromatogram signal for quantifier, qualifier and internal standard transitions for both analytes.

4.3.2 Reorganizing the collected data

Processing of the files produced two different arrays including data from the measurements. This was required to form similar QC sets that were used in the RSD studies with the current integration method. Processing the raw files with the Python script did not retrieve channel and analyzer information of the measurements. That information was required to form the measurement sets. The required measurement sets could be formed by combining the two arrays by connecting the channel and analyzer information to each individual ion measurement based on the unique JobID and analyte name. In total the channel and analyzer information was linked to 7200 measurements, forming an array including 120 quality control sets, 30 for each channel and analyte combination. Each quality control set had 20 measurements which all included chromatograms for the three different transitions. The combination of the arrays were performed with Python, with the usage of pandas package.

From the formed array it is possible to retrieve the chromatograms of certain QC sets. The proposed peak detection and integration method could then be applied to these chromatograms, as well as the existing methods from the literature. The RSD of the QC set could then be compared to the RSD obtained with the current method. The RSDs of the current method were stored in separate array from which each individual measurement set could be identified based on analyzer, analyte and measurement channel and compared to the RSD obtained with the other methods.

4.4 Peak detection and integration pipeline

Peak detection and integration pipeline was developed as part of this study to detect and integrate the obtained chromatogram peaks. The purpose of the pipeline was to detect and integrate peaks with defined methods in the pipeline. Multiple different methods were applied to the pipeline. The combined data in data acquisition phase was used as an input for the pipeline. The pipeline was developed with Python using pandas, NumPy and Scipy packages.

4.4.1 Data processing

The data processing starts by giving the array formed in the data acquisition phase as input to the peak detection and integration pipeline. In addition to this, a list including the order in which the excel files from previous RSD studies in data acquisition phase were processed was given as input. This was done in order that the RSD results from previous studies could be compared easily to the calculated RSDs with the proposed methods. In the beginning of the pipeline the given input is divided to four different arrays. This is done based on the analyte name and the channel used in the analyzer for the measurement. Four arrays cover all the different combinations that two analytes and two channels can form. E.g. one array includes all measurements with channel 1 and analyte 1.

After the input from data acquisition phase is divided to four sets based on analyte and analyzer channel, the intensities of the chromatogram and the measurement times of each transition (quantifier of analyte x , qualifier of analyte x and internal standard) are saved to separate arrays forming 6 different arrays. This is done with a function looping through the input array (one of the four arrays formed in previous phase of data processing). In the looping process the intensities and measurement times of an individual transition are fetched from the input array and saved to one of the six arrays depending on the transition. In order to check that the values saved to the arrays are in the correct order (e.g. index i in array including quantifier intensities is from the same measurement as index i in array including qualifier intensities) checking of JobIDs of each individual measurement is done. One unique JobID includes one measurement with each transition. That is why by checking the JobID in each iteration makes sure that the values are saved in correct order. The values must be in the correct order in order to be able to calculate the responses correctly in later phase.

These six arrays formed can then be given to peak detection and integration function. The peak detection and integration function is discussed in more detail in the next chapters. This is done in main function of the pipeline. In the main function, the input data is first divided to four arrays based on the analyte and channel. Then in the main loop, each of these arrays is given to a function which processes the 6 different arrays including measurement times and intensities of each transition one by one. The correct order for the analyzers is checked from the list that is given as input. In each iteration, the peak integration function returns the calculated areas of fitted peaks or the real chromatogram signal depending on the method used, for the quantifier and internal standard transitions for each of the 4 quality control test set

for one analyzer. From these values the RSD can be calculated for each set. Then these values are saved for each analyzer and analyte/channel combination giving us 120 RSD from 30 different instruments. In the result analysis phase, saved values are compared to the values given in excel files of previous RSD studies which match to the same exact measurements.

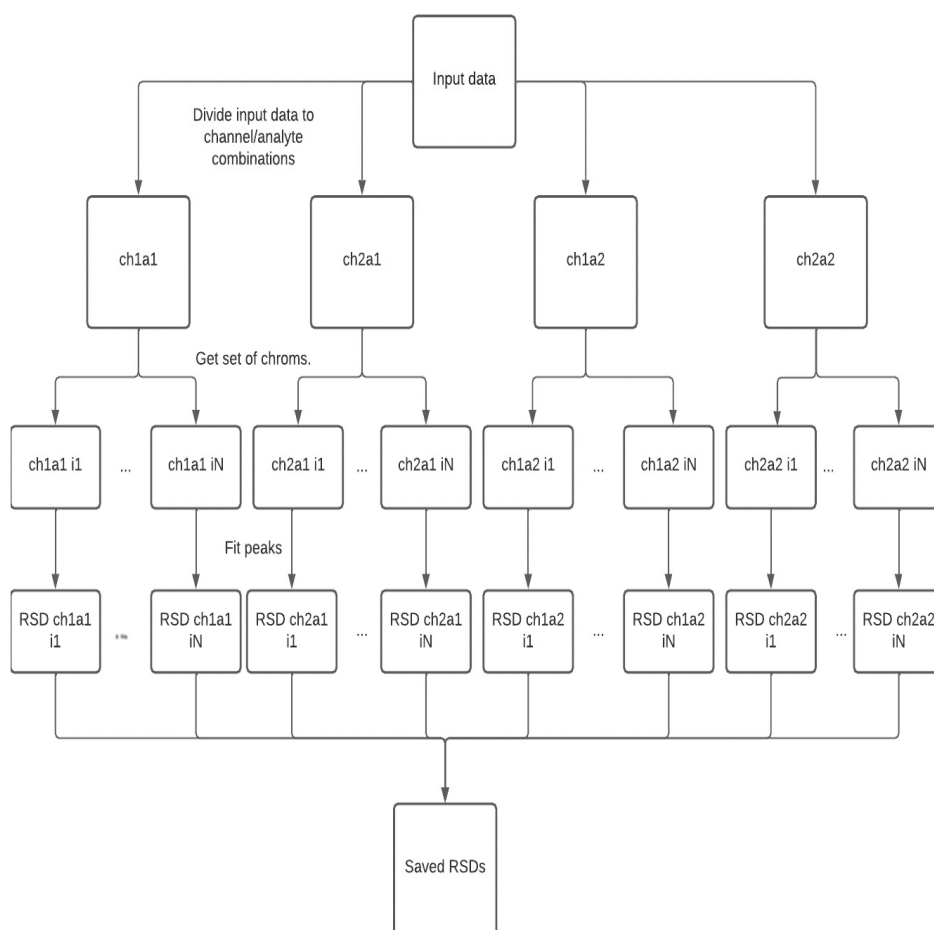


Figure 5: Data processing of chromatograms illustrated. The input data is first divided to different channel and analyte combinations. Then from each combination chromatograms are obtained for each instrument forming 30 arrays for each combination. Then from these array the RSDs of the responses are calculated and saved for further investigation. Abbreviation *ch* presents the channel, *a* presents the analyte and *i* presents the instrument (analyzer)

4.4.2 Peak integration function

A separate function was developed for experimenting with different peak detection and integration methods. The input for the function was given as a set of intensities and measurement times. The intensities and measurement times were already divided

for each transition, providing 6 different arrays (3 intensity arrays and 3 measurement time arrays). Each of the arrays have 20 intensity value or measurement time sets with the length of 160. The functionality of the function starts by looping through the input arrays. In each iteration the chromatogram of internal standard, including 160 intensities, is selected. From this chromatogram the range including peak start and end points of the peak are estimated by finding the maximum value of the chromatogram. The signal to noise ratio with internal standard is so good that it can be expected that the maximum value of the chromatogram is always part of the top of the peak. From this point, a range of 50 intensities and measurement times is selected. The maximum point is in the middle of this range. This range always includes the peak, and this method also reduces a significant part of the noise.

A similar process is done for quantifier and qualifier chromatograms. However, in order to get the quantifier and qualifier peaks correctly every time, a range of intensities was selected from the chromatogram based on the index in measurement times of internal standard peak. The internal standard peak has significantly less noise compared to the quantifier and qualifier peaks, because the internal standard peak is much stronger peak, i.e. the intensity values are much greater. The index in measurement times of internal standard peak can be used to estimate the index in measurement times of quantifier and qualifier peaks, because the three transitions are measured together and we have prior knowledge that the peaks are formed approximately at the same time in the chromatogram. From the maximum index of internal standard peak, the same number of intensities and measurement times were taken from the quantifier and qualifier transitions, as previously done for the internal standard. The range was large enough so that the peak was included for both transitions in all measurements. This means that some noise remained in the chromatograms but significant amount of noise was reduced from the start and end points of the chromatogram. After obtaining the peaks for integration from the chromatograms, different peak integration methods could be applied. The method for integration could be varied inside the function.

4.4.3 Integrating peaks based on fitted theoretical peaks

In this section a peak integration method based on fitting a theoretical peak to the chromatogram data is presented. The peak fitting was based on nonlinear least-squared fitting described in chapter 3.5. Levenberg-Marquardt method of the least-squared fitting was used in the fitting. An existing implementation of this, `scipy.optimize.least_squares` provided by SciPy was used.

The intensity value ranges and measurement times acquired from the first part of the peak integration function were then used for fitting of the peaks. The fitting was performed by combining the three separate peak fittings. The combined fitting method was believed to reduce the effect caused by the noise, because the shape of the three chromatogram peaks from the same measurement should be the same because they are from the same analyte. That lead to the development of combined fitting in this thesis.

As prior knowledge, we knew that the quantifier and qualifier ion peaks should be

in the same exact locations on the measurement time axis. It was also know that the peak location for internal standard is close to the quantifier and qualifier peaks but slightly different. In addition to this, it was know that all the peaks should have the same standard deviation. From this information a combined nonlinear least squares fitting problem was formed. Two different theoretical functions were selected to be fitted, the Gaussian function 5 and the EMG function 8.

When a function was selected for optimization, all the three peaks included some shared parameters. For example, the combined fit for the Gaussian function 5 is given by

$$\begin{aligned} \chi^2(a1, a2, a3, b1, b2, c) = & \sum_{i=1}^m \sqrt{w1_i} [(y(t_{1i}) - \hat{y}(t_{1i}; a1, b1, c))]^2 + \\ & \sum_{i=1}^m \sqrt{w2_i} [(y(t_{2i}) - \hat{y}(t_{2i}; a2, b1, c))]^2 + \\ & \sum_{i=1}^m \sqrt{w3_i} [(y(t_{3i}) - \hat{y}(t_{3i}; a3, b2, c))]^2, \end{aligned} \quad (15)$$

where the parameters $a1$, $a2$ and $a3$ where the amplitudes for quantifier, qualifier and internal standard peaks in this order. $b1$ was the shared location of the peak center for quantifier and qualifier peaks which form the prior knowledge should be the same. $b2$ was the location for the peak center for the internal standard peak and c was the standard deviation of a peak, which from the prior knowledge should be the same for all peaks. $t1$, $t2$, $t3$ are the measurement times for quantifier, qualifier and internal standard, and $w1$, $w2$, $w3$ are the weights for these transitions. The function \hat{y} optimized is the theoretical peak selected for combined fitting.

The intensity values were assigned as weights for the functions in the combined fitting. It was desired that the high intensity values at top of the peak got more weight in the fitting than the noise in the data at start and end of the chromatogram. The weight at each point i was calculated from the chromatogram intensities with formula

$$w_i = \frac{\sqrt{chrom_n(i) + 1}}{chrom_n(i) + 1}, \quad (16)$$

where $chrom_n$ is the given chromatogram. One is added to the intensities in case of intensity value happens to be zero.

In the case of fitting the EMG 8 peak, one extra parameter was added to the functions optimized. This parameter represented the skewness of the peaks. This parameter was added to all the three functions in the combined fitting. The skewness parameter was handled as an individual parameter for all the peaks. The optimization returned optimal values for each peaks parameters which formed the three theoretical peaks based on the combined fitting and function selected for the fitting. After obtaining the three theoretical peaks, they could be decided to represent the real peaks or the integration could be performed based on the location of the theoretical peaks. Integrating the real peak based on the location of the fitted peak was performed so

that a threshold intensity value was decided for integration limits. The integration was performed from time points when the fitted peak intensity value matched this threshold value, and the real peak was integrated from these points. The threshold value was a certain percentage of the top intensity value of the real peak. The exact percentage is presented in the chapter describing the execution of the peak integration function. Fitting of both theoretical peaks was performed and RSDs calculated from the result integrated areas which is discussed in later chapters.

4.4.4 Adding baseline and CWT method

A baseline approximation option was added as part of the combined fitting method. The combined fitting method remained otherwise the same but the baseline from each chromatogram was approximated and subtracted from the chromatogram before performing the combined fitting. Two different baseline functions were tested separately with the combined fitting, SNIP baseline estimation function 14 and Beads baseline estimation function described in chapter 3.4. A ready implementation of Beads was used from PyBeads Python package. The option for using one of the baseline estimation function from the literature was added as part of the peak integration function entity. Both of these baseline estimation methods were experimented for integration of the real peaks and the fitted peaks of the Gaussian function.

By varying the baseline option, fitted theoretical peak option (Gaussian or EMG) and integration peaks (real or fitted) multiple different results could be produced from the QC data. These proposed method entities should also be compared to some existing method from the literature in addition to the current method used in the analyzers. Continuous wavelet method described in chapter 3.2 was selected for this method. The CWT method was added as part of the peak integration function, and the combined fitting function was replaced with the CWT method when the CWT method was decided to be used. The points for integrating the real peaks were decided based on zero crossing points of the CWT approximation. The CWT method was executed with Scipy.signal.cwt Python package and the selected scale for the CWT was obtained based on the full width at half maximum. The scale was calculated from formula

$$s = (2 * \sqrt{2 * \log 2 * std})/2,$$

where *std* is the standard deviation of the peak. The result of the function was rounded to closest integer.

The CWT method uses second derivative of Gaussian to approximate the start and end points of the chromatographic peaks Du et al. [22]. This differs from the proposed method in this thesis. The proposed method fits Gaussian or EMG peak to the data instead of using derivative method. Then the integration points of the real peaks are approximated by integrating the part of the peak which is above the decided threshold. Also, the CWT method differs from integrating the fitted peak because the CWT is used to approximate the start and end point of the real peak, instead of integrating the fitted peak. In XCMS [4] no prior baseline estimation is

performed before usage of the CWT method. No baseline estimation function was used before using the CWT method in these experiments.

4.5 Execution of the methods

The proposed combined peak fitting method with Gaussian and EMG functions, baseline approximation prior the combined fitting and CWT method were experimented with the QC data sets. The intensity value and measurement time ranges of quantifier, qualifier and internal standard were given as input to the peak integration function to perform these methods. The optimized result parameters were used to form new fitted peaks for the quantifier and internal standard. These fitted peaks were integrated and they were also used to determine integration limits for the real peaks. The purpose of the combined fitting was to improve the RSD of responses from the quality control sets. The response for one measurement was calculated by dividing the integrated area of the quantifier peak with the integrated area of the internal standard peak obtained from the methods. It can be described with formula

$$r = QA/IA,$$

where r is the response, QA is the quantifier peak area and IA is the internal standard peak area. The advantage from this method was that prior knowledge of similarity of the three peaks could be used to form the fitted quantifier and internal standard peaks. The fitted peaks are expected to have less noise compared to the chromatogram data because some the peak parameters are optimized from three peaks. The functionality of the peak fitting function is be described with pseudo code in 6:

The usage of the peak integration function required definition of initial guess parameters for the combined fitting, threshold intensity value integrating the real peaks based on fitted peaks and parameters for the baseline estimation functions. For initial guesses for the combined fitting the maximum value of the chromatogram was taken as initial guess for the amplitude, the measurement time of the maximum value was taken as initial guess for the position of the center of the peak and standard deviation guess was set to 0.035. The standard deviation guess was updated in each iteration. In case of the EMG function, initial guess for the skewness of the peaks were varied between 0.02 – 0.04.

The threshold for integration of real peaks based on Gaussian fitting was set so that the range for integration was the time range where the fitted Gaussian peak intensity were over 7.5% of the peak maximum. For EMG function the percentage was set to 10% because the remained noise in the chromatogram seemed to effect the tailing of the peak. The fitted peaks were integrated from the whole range of the time axis of the chromatogram. If a baseline approximation method was applied to the chromatogram data prior combined fitting, the threshold value for integration was set to 4%. The chromatograms were still expected to have some noise after baseline approximation. For SNIP baseline estimation function, the number of iterations performed to obtain the baseline estimation was set to 15. The same number of iterations were performed with the Beads baseline estimation algorithm. Initial

```

function PeakIntegration(quant_chroms, qual_chroms, is_chroms, quant_t, qual_t, is_t, std0, threshold)
    quant_fitted_areas = []
    quant_real_areas = []
    is_fitted_areas = []
    is_real_areas = []
    for i in the range [0, len(quant_chroms)]
        ISmaxi = MaxIndex(is_chroms[i])
        isT, isC = is_t[i][ISmaxi-25:ISmaxi+25], is_chroms[i][ISmaxi-25:ISmaxi+25]

        quantT, quantC = quant_t[i][ISmaxi-25:ISmaxi+25], quant_chroms[i][ISmaxi-25:ISmaxi+25]

        qualT, qualC = qual_t[i][ISmaxi-25:ISmaxi+25], qual_chroms[i][ISmaxi-25:ISmaxi+25]

        if estimate_baseline
            isC = isC - baseline(isC)
            quantC = quantC - baseline(quantC)
            qualC = qualC - baseline(qualC)

        initial_guess = [quant_chroms[i][ISmaxi], qual_chroms[i][ISmaxi], is_chroms[i][ISmaxi],
            mean([quant_t[i][ISmaxi], qual_t[i][ISmaxi]]), is_t[i][ISmaxi], std0]

        optimized_parameters = least_squares(combined_fitting, initial_guess, arguments)

        quant_args = optimized_parameters[0]
        is_args = optimized_parameters[1]

        fitted_quant_peak = peak_function(quant_args)
        fitted_is_peak = peak_function(is_args)

        quant_fitted_areas[i] = integrate(fitted_quant_peak)
        is_fitted_areas[i] = integrate(fitted_is_peak)

        quant_integration_limits = get_integration_limist(fitted_quant_peak, threshold)
        is_integration_limits = get_integration_limist(fitted_is_peak, threshold)

        quant_real_areas[i] = integrate(quantC, quant_integration_limits)
        is_real_areas[i] = integrate(isC, is_integration_limits)

    return quant_fitted_areas, quant_real_areas, is_fitted_areas, is_real_areas

```

Figure 6: Pseudo representation of the peak integration function as whole.

guesses of sparsity of the signal and its derivatives were calculated by checking the number of points above estimated noise level.

5 Experiments and results

Multiple experiments were performed with the framework described in previous chapter. In this chapter, results of integrating the real peak versus the fitted peak are compared, the effect of the channel is investigated, the possible effect of baseline estimation prior fitting peaks is investigated and the results from different peak integration methods are compared.

5.1 Integration of fitted and real peak

In this section, the results from integrating the fitted peak are compared to the results of integrating the real peak based on the fitted peak and threshold value for integration. Result data from the usage of Gaussian curve and EMG curve without the usage of baseline estimation are compared here. The results were produced with the framework described in previous chapter. Combined 120 RSDs were obtained from the framework for each method, 30 for each channel analyte combination. Each individual RSD was calculated from 20 responses.

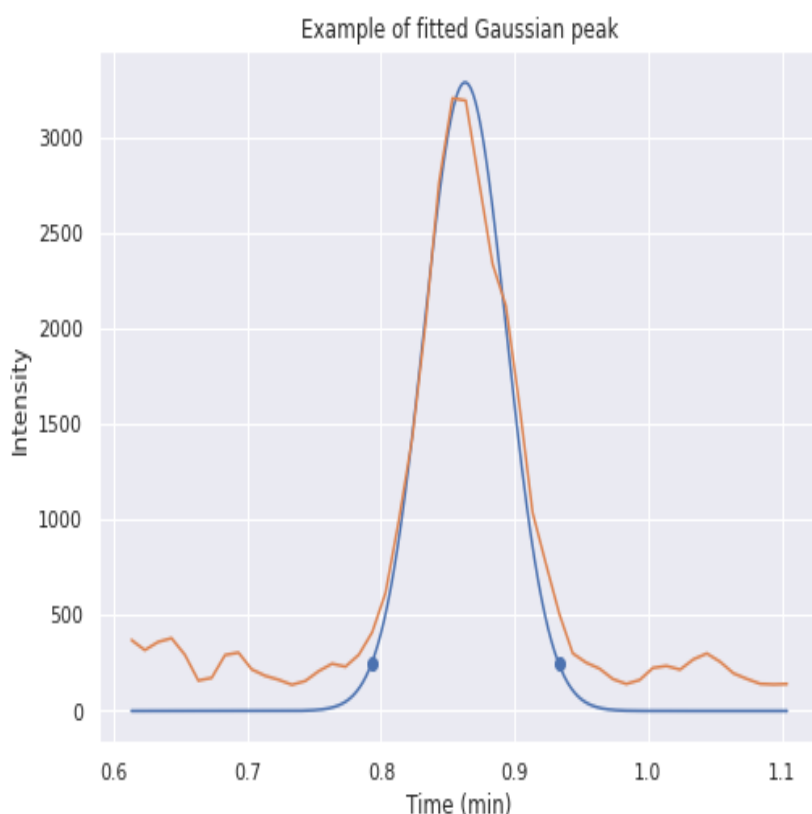


Figure 7: Example of fitted Gaussian peak to analyte signal. The analyte signal is plotted as orange and the fitted Gaussian as blue. The blue points represent the threshold points for integration of the real peak. The analyte signal is from quantifier transition measurement.

Table 1: Integrating the fitted peak versus the real peak is compared in Table 1. Peak column includes the fitted peak function, Analyte column includes the analyte type, Ch column includes the used channel in measurements, P and T- value columns include result values from comparisons, FM column includes fitted method mean of RSD percentages and RM column includes real peak method mean of RSD percentages

Peak	Analyte	Ch	P-value	T-value	FM	RM
Gaussian	Analyte1	1	0.530	0.63	5.49	5.29
Gaussian	Analyte1	2	0.663	0.44	5.46	5.35
Gaussian	Analyte2	1	0.063	1.91	4.32	3.89
Gaussian	Analyte2	2	0.076	1.81	4.74	4.19
EMG	Analyte1	1	0.024	-2.33	5.22	5.95
EMG	Analyte1	2	0.001	-3.39	5.30	6.35
EMG	Analyte2	1	0.000	-7.11	4.14	6.82
EMG	Analyte2	2	0.000	-5.56	4.76	7.03

For the comparison of integrating fitted peak versus real peak, two sample t-test with 95% confidence level was performed for each fitted versus real peak pair. Here, a pair refers to the RSDs obtained with the same function in the combined fitting and the same analyte channel combination in QC measurement sets. The hypothesized difference was set to zero in the two sample t-tests. The results are presented in the following table and they were obtained with Minitab software.

From the table 1, it can be observed that the mean RSD percentage is lower for integrating the real peak in case of fitting the Gaussian peak. However, it can not be observed with statistical significance that integrating the real peak produces lower RSDs because the obtained p-values from the two sample t-tests are all over the 0.05 significance level. It can also be observed that the significance level is close to the 0.05 significance level in case of analyte 2, and that there is a difference in RSD percentage means between the methods.

In the case of fitting the EMG peak, integrating the fitted EMG peak produces significantly lower RSD results. This can be observed from the low p-values from table 1. The difference of RSD means between the two methods is high for both analytes. This is quite expected result because the EMG peak is assymmetric. Integrating the real peak based on the intensity value on the fitted peak could cause variance among the integration limits among the peaks because the tailing among the peaks might vary. This might be the cause for higher RSDs when integrating the real peak based on the fitted EMG peak. However, this is not the case with Gaussian peak because it is a symmetric peak. For Gaussian peak, integrating the real peak when fitted Gaussian intensity is over a certain threshold is more suitable because tailing is not present. It can be also noticed that integrating the fitted EMG peak seems to produce slightly lower RSDs than fitted the Gaussian peak. From these results, integrating fitted EMG peak and integrating the real peak based on Gaussian fitting methods are selected for further comparison against other methods.

The following results are also represented with interval plot below in figure 8.

In the interval plot, the intervals should be compared as pairs. The interval plot represents all the analyte and channel combinations for integrating the fitted peak and the real peak based on the fitted peak. A1/A2 is abbreviation for the analyte, Ch1/Ch2 is abbreviation for the channel F/R is abbreviation for fitted and real peak and G/E is abbreviation for the Gaussian and EMG peaks.

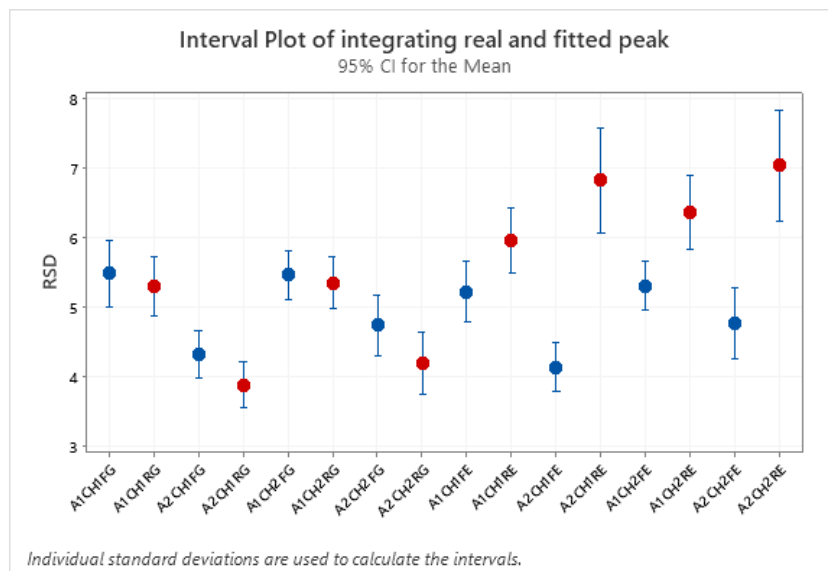


Figure 8: Interval plot of integrating the fitted peak versus the real peak based on the fitted peak pairs. Results obtained by integrating the fitted peaks are presented as blue and integrating real peaks are presented as red.

5.2 Channel and baseline effect to results

In this chapter the effect of used channel to results is compared. Also, the effect of subtracting the baseline prior peak fitting method is investigated. The used channel in the analysis with the analyzers should not have effect to the results. The goal is that both channels produce reliable results that can be used and analyzed as one set. However, the channel is a variable in the analysis process which is why its effect is analysed here.

The results between channels were analysed by comparing the results obtained from the currently used method and integrating the real peak based on the fitted Gaussian function. The results between channels were compared with two sample t-tests. The results are presented in the following table and they were obtained with Minitab software.

Table 2: The effect of channel is compared in Table 2. Method column includes the used method to obtain the RSDs, Analyte column includes the analyte type, P and T value columns include result values from comparisons, CH1M column includes mean of RSD percentages in channel 1 and CH2M column includes mean of RSD percentages in channel 2.

Method	Analyte	P-value	T-value	CH1M	CH2M
Gaussian	Analyte1	0.833	-0.21	5.29	5.35
Gaussian	Analyte2	0.266	-1.21	3.89	4.19
Current	Analyte1	0.946	-0.07	6.80	6.83
Current	Analyte2	0.263	-0.263	6.43	6.91

From the table 2, it can be observed that the channel does not cause statistically significant difference in results. The observed p-values are high for both analytes, thus it can be observed that the difference between channels is greater with analyte 2 with both methods. Because the channels do not have a significant effect on the results it can be argued that the results can be compared only based on analyte and method.

Two different baseline approximation methods were performed prior the fitting of Gaussian peaks. These results from the usage of these baseline methods were compared to the results obtained without baseline estimation. The used baseline estimation functions were the SNIP baseline estimation [34] and BEADS baseline estimation [35]. The baseline estimation methods were compared to the combined fitting without baseline estimation with two sample t-tests. Integrated real peaks based on combined fitting were compared for the Gaussian peak because it produced the lowest average mean of RSDs for both analytes in previous chapter. The results can be observed from table below.

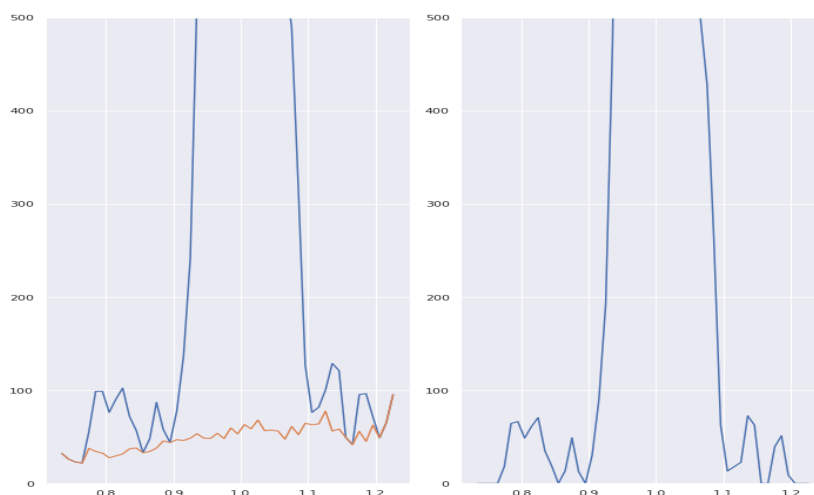


Figure 9: Example of SNIP baseline estimation. On the left side, the chromatogram and estimated baseline are plotted. On the right side, the chromatogram is plotted after baseline is subtracted from the chromatogram.

Table 3: Results obtained with adding baseline estimation to the method are compared against the results without baseline estimation in Table 3. Method column includes the used method to obtain the baseline, Analyte column includes the analyte type, P and T value columns include result values from comparisons, BM column includes mean of RSD percentages with the baseline method and NM column includes mean of RSD percentages without baseline method.

Method	Analyte	P-value	T-value	BM	NM
SNIP	Analyte1	0.960	0.05	5.31	5.32
SNIP	Analyte2	0.000	-4.76	4.99	4.04
BEADS	Analyte1	0.252	-1.15	5.55	5.32
BEADS	Analyte2	0.404	-0.84	4.20	4.04

From the table 2, it can be observed that estimating the baseline and performing combined fitting afterwards do not improve the results. With both baseline estimation methods, the average of RSD percentages are close to each other with analyte 1. The p-value of SNIP method comparison suggest that the SNIP method does not affect the results at all in case of analyte 1. In case of the analyte 2, it can be observed that adding the SNIP method actually makes the obtained results worse because the p-value is below the 95% significance level. In case of using BEADS baseline estimation, no statistically significant effect can be observed based on the p-values but it can be noticed that with both analytes the usage of BEADS slightly increased the average RSDs.

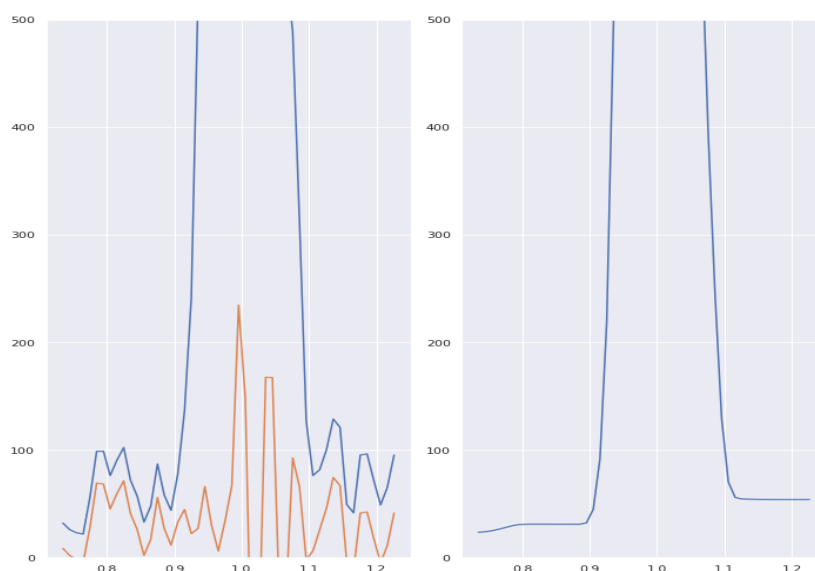


Figure 10: Example of BEADS baseline estimation. On the left side, the chromatogram and estimated baseline are plotted. On the right side, the chromatogram is plotted after baseline is subtracted from the chromatogram.

The figure 10 show how the algorithm is able to estimate the baseline and the noise. It can be observed that the BEADS estimation algorithm is able to estimate

noise peaks quite well and the signal becomes quite smooth after estimation. However, subtracting the noise from the signal did not improve the obtained RSDs. From figure 9, it can be observed that the SNIP algorithm estimates the baseline quite well but the signal still includes noise peaks. The SNIP algorithm also did not improve the obtained RSDs.

5.3 Comparison of the methods

In this section, the proposed method is compared against the currently used method in the analyzers and the CWT method. Integrating fitted EMG peak and integrating real peak based on fitted Gaussian peak were chosen to the comparison. Baseline estimation was not included in these methods. The CWT method was experimented with the same peak detection and integration pipeline by replacing the combined peak fitting method with the CWT method. RSDs from the QC runs with the current method were obtained from processing the excel result files from these runs.

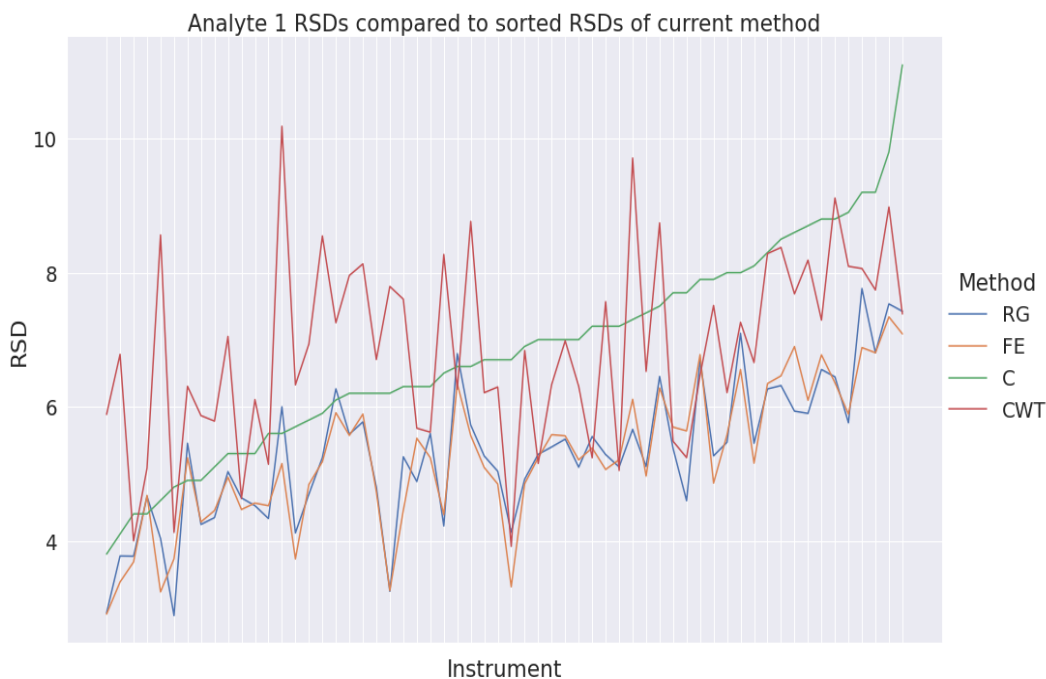


Figure 11: Analyte 1 RSDs of each method. The current method RSD values are sorted to make the visualization more clear. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method. Horizontal axis includes used analyzers to produce the QC results.

Table 4: Results of peak integration methods for analyte 1 compared with two sample t-tests. Method1 and method2 columns include the methods in comparison, P and T value columns include result values from comparisons, Mean1 column includes mean of RSD percentages for method 1 and Mean2 column includes mean of RSD percentages for method 2 in comparison. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method.

Method1	Method2	P-value	T-value	Mean1	Mean2
RG	C	0.000	-6.25	5.32	6.82
RG	CWT	0.000	-6.77	5.32	6.87
FE	C	0.000	-6.51	5.26	6.82
FE	CWT	0.000	-7.05	5.26	6.87
FE	RG	0.763	-0.30	5.26	5.32
C	CWT	0.841	-0.20	6.82	6.87

From figure 11, it can be observed that the integration of real peak based on fitted Gaussian and integration of fitted EMG peak seem to produce lower RSD percentages than the currently used method and the CWT method. It also seems that the current method might have lower mean of RSD values than the CWT method, thus CWT outperforms the current method at some points with the analyte 1. Clear difference between integrating fitted EMG and real peak based on fitted Gaussian can not be observed from the figure 11. These observations were investigated further with two sample t-tests and interval plot.

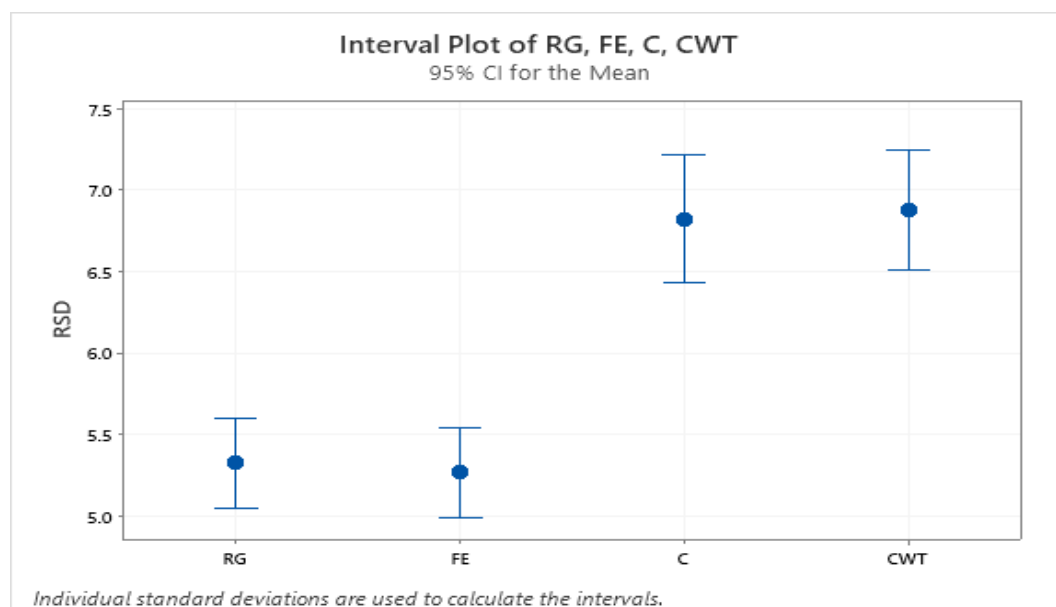


Figure 12: Interval plot of different methods with analyte 1. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method.

From table 4 and interval plot 12 it can be observed that integrating the real peak based on Gaussian fitting and integrating fitted EMG peak methods outperform the current and CWT methods. The p-values of two sample t-tests show statistically significant difference when these methods are compared. Also, the mean difference of RSDs between proposed methods and the two methods compared differs over 1 percent unit which is quite significant improvement. Statistically significant difference can not be observed from the results when the two proposed methods are compared to each other or when the current method is compared to the CWT method. In these comparisons, the RSD means are close to each other.

The obtained results were compared further with a correlation plot. In the following correlation plot the difference between RSDs of current method and integrating the real peak based on Gaussian fitting is compared to the RSDs of the current method. The goal of this comparison was to see if the proposed method is able to improve the RSD more if the RSD with the current method is high than when the RSD with current method is low. The difference is calculated by subtracting RSDs obtained with the proposed integration method from the RSDs obtained with the currently used method.

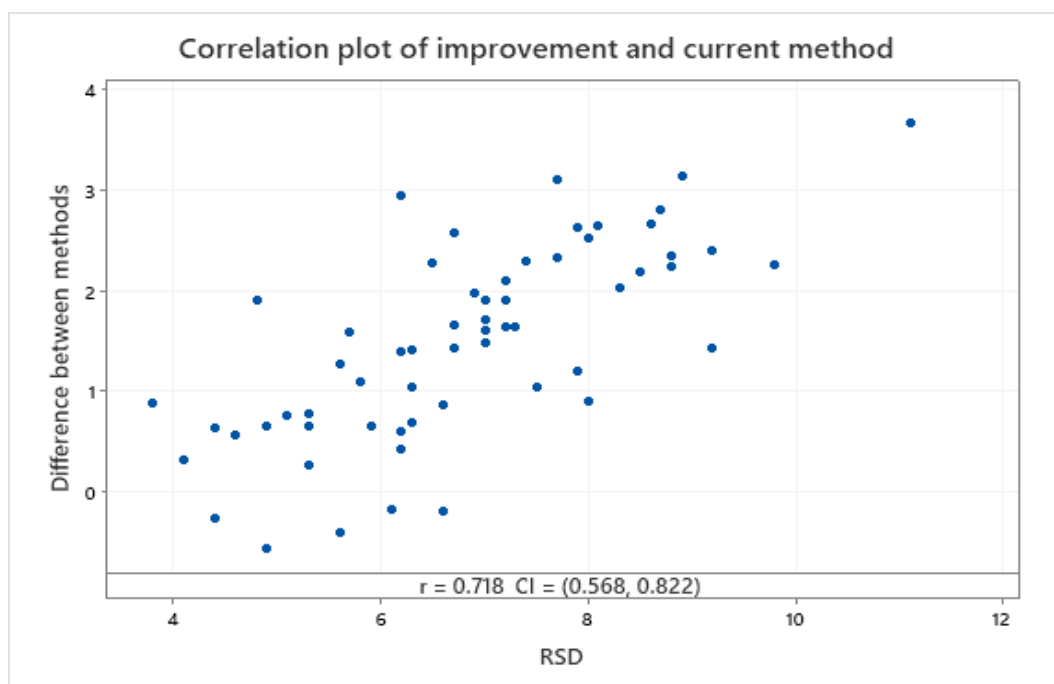


Figure 13: Correlation plot of analyte 1. The plot represent correlation between the difference between current method and integrating the real peak based on fitted Gaussian and the current method. The difference is calculated by subtracting the RSD of fitted Gaussian method from the current method. It represents the improvement obtained by the proposed method.

The correlation comparison in figure 13 was performed with Pearson correlation with 95% confidence level. The obtained Pearson correlation coefficient and confidence interval have quite high values, which indicate positive correlation between the RSD

differences of the two methods and RSDs of the current method. If the threshold for moderate correlation is set to 0.5, moderate positive correlation can be observed because the confidence interval is over this threshold value.

The next step in the result analysis was to perform similar analysis with the analyte 2. Results were produced with all the same methods that were used to produce results for analyte 1. The obtained results are visualised in the figure below.

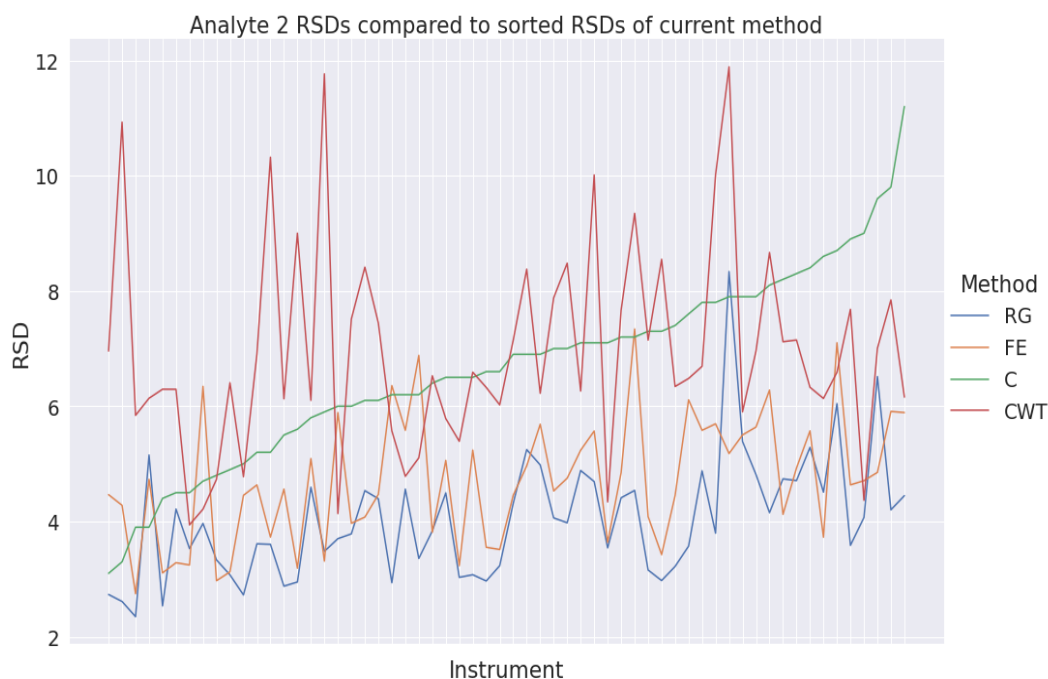


Figure 14: Analyte 2 RSDs of each method. The current method RSD values are sorted to make the visualization more clear. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method. Horizontal axis includes used analyzers to produce the QC results.

From figure 14, it can be observed that the integration of real peak based on fitted Gaussian and integration of fitted EMG peak seem to again produce lower RSD percentages than the currently used method and the CWT method. With analyte 2, there also seems to be a difference between the RSDs of the two proposed methods. In the case of analyte 1, a difference in the methods could not be observed as clearly. The CWT method seem to produce a little bit higher results with analyte 2 than analyte 1. These observations from the visualization of RSDs were investigated further with two sample t-tests.

Table 5: Results of peak integration methods for analyte 2 compared with two sample t-tests. Method1 and method2 columns include the methods in comparison, P and T value columns include result values from comparisons, Mean1 column includes mean of RSD percentages for method 1 and Mean2 column includes mean of RSD percentages for method 2 in comparison. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method.

Method1	Method2	P-value	T-value	Mean1	Mean2
RG	C	0.000	-10.52	4.04	6.67
RG	CWT	0.000	-10.83	4.04	6.95
FE	C	0.000	-7.68	4.72	6.67
FE	CWT	0.000	-8.18	4.72	6.95
FE	RG	0.001	3.47	4.72	4.04
C	CWT	0.370	-0.90	6.67	6.95

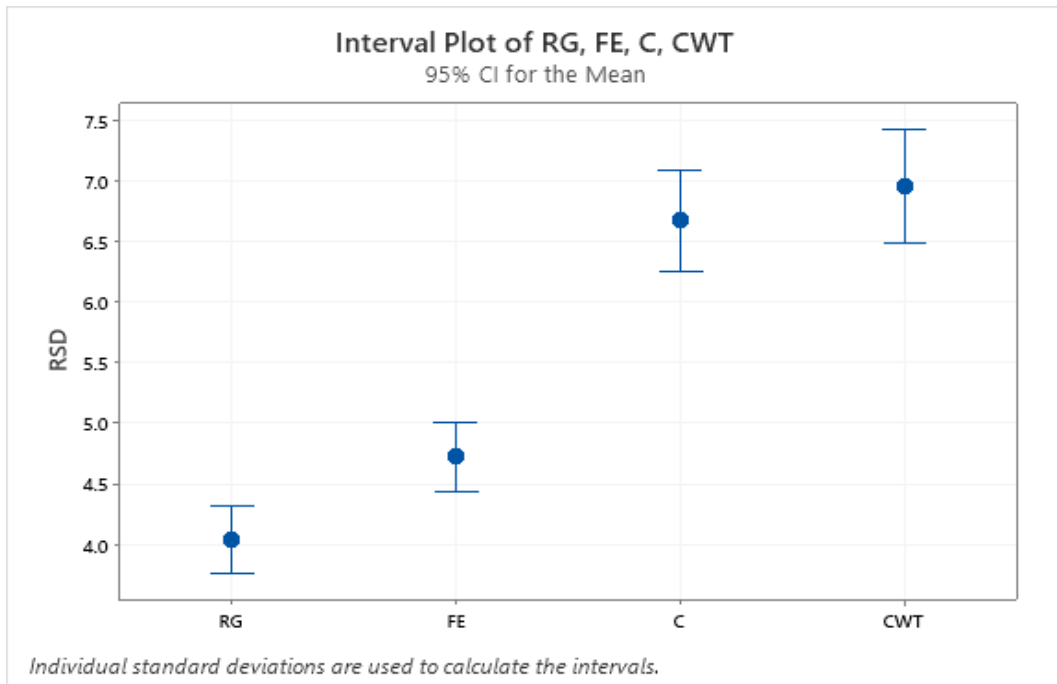


Figure 15: Interval plot of different methods with analyte 2. RG refers to the integration of real peak based on Gaussian fit, FE refers to the integration of fitted EMG peak, C refers to the current method and CWT refers to the CWT method.

From table 5 and interval plot 15 it can be observed that integrating the real peak based on Gaussian fitting and integrating fitted EMG peak methods outperform the current and CWT methods. Also, it can be observed that the method integrating the real peak based on Gaussian fitting outperform the method integrating the fitted EMG peak. This observation could not be made in case of analyte 1. The p-values of two sample t-tests are statistically significant in these comparisons. When comparing

the average RSDs, the difference between integrating the real peak based on Gaussian fitting and the current and CWT method is even greater than with analyte 1. The current method produced a little bit lower average of RSDs than the CWT method but statistically significant difference between these two method can not be observed.

The obtained results were again compared further with correlation plot. In the following correlation plot the difference between RSDs of current method and integrating the real peak based on Gaussian fitting is compared to the RSDs of the current method. The difference is calculated by subtracting RSDs obtained with the proposed integration method from the RSDs obtained with the currently used method.

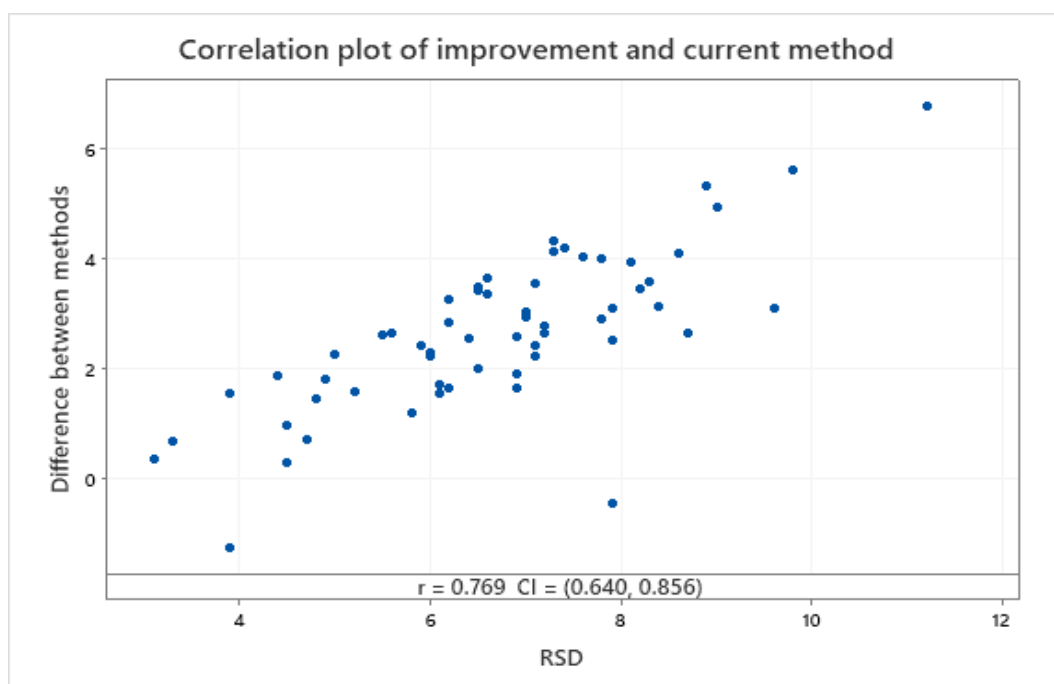


Figure 16: Correlation plot of analyte 2. The plot represent correlation between the difference between current method and integrating the real peak based on fitted Gaussian and the current method. The difference is calculated by subtracting the RSD of fitted Gaussian method from the current method. It represents the improvement obtained by the proposed method.

The correlation comparison in figure 16 was performed with Pearson correlation with 95% confidence level. The obtained Pearson correlation coefficient and confidence interval have even higher values values than with analyte 1. That is why moderate positive correlation between the RSD differences of the two methods and RSDs of the current method can be observed. The confidence interval values are over the 0.5 threshold which indicates the moderate positive correlation. A couple of outlier points can be observed from both correlation plots 13 16. In these cases the current method produces lower RSDs. The outliers could be caused by certain type of peak shapes that the proposed methods are not able to represent well. However, this does not seem to be common among the QC sets. But overall, it can be observed that

the proposed method produce lower RSDs for the QC sets than the currently used method among the analyzers. Also, the improvement seem to be greater when the current method produce high RSD percentages.

6 Conclusion

Various different peak detection and integration methods have been developed in order to produce precise results from LC-MS analysis method. Keeping up a good precision usually becomes more difficult when the s/n ratio of the signals are low. Then the noise causes variation in determining the start and end points of the peak, even false positive peaks might be detected. In order to produce precise results, the used peak detection and integration algorithm needs to be able to handle noise in the signal.

There are many open source software packages to process raw files and perform integration of chromatograms produced with a LC-MS analysis method. A new framework developed to process raw files and reorganize the obtained data in a specific way was developed as part of this thesis. The peak detection and integration algorithm developed as part of this thesis was successfully implemented as part of the framework. The framework was able to process raw files and save the desired data from the files for further investigation. The data was reorganized as part of the framework in a desired way to apply peak integration methods in QC sets. Multiple peak integration and baseline estimation combinations could be performed with the developed framework.

The proposed peak detection and integration algorithm was based on fitting a theoretical peak to the chromatograms. The peak fitting was performed as combined fitting, where the same theoretical peak was fitted to the quantifier, qualifier and internal standard chromatogram peaks formed by their transitions. The fittings were performed simultaneously in the same optimization function. The idea of simultaneous fitting was based on the fact that quantifier, qualifier and internal standard shared some peak parameters, which in theory should be the same. This lead to the development of combined peak fitting.

In addition to integrating the fitted peaks, the real peaks were integrated based on the fitted peaks. The real peaks were integrated in the range when the fitted peak intensity value matched a defined threshold value. The threshold value was set based on the top value of the fitted peak. The results showed that in the case of fitting the Gaussian peak, integrating the real peak based on the fitted Gaussian produced lower RSD averages than integrating the fitted Gaussian peak. However, the RSDs were close to each other with both methods and difference in the methods could not be observed with statistical significance. With analyte 2, the p-values were close to the 95% significance. In the case of fitting the EMG peak, integrating the fitted peak instead of integrating the real peak based on the fitted peak proved to be a better method. There was a major difference in average RSDs and the difference in RSDs was statistically significant according to p-values obtained in two sample t-tests. Baseline estimation prior the combined fitting was experimented with, but it did not improve the results of combined fitting. This conclusion was also made based on p-values of two sample t-tests. The result RSDs obtained without baseline estimation were compared against result RSDs obtained with baseline methods added to the combined fitting method.

The proposed peak detection and integration methods developed performed well

compared against the CWT method from the literature and the current method in the analyzers from which the data was obtained for this thesis. Two different proposed methods were used in the comparison which were the method integrating real peak based on fitted Gaussian and integrating the fitted EMG. The proposed methods were able to improve the RSD in most of the QC sets with both analytes. When the method integrating the real peak based on fitted Gaussian was compared against the current method, the RSD of a QC set improved the most when the current method produced high RSD percentages. This conclusion was made based on correlation study. The high RSD percentages obtained with the current method could be the cause of significant amount of noise in the chromatograms. That could cause the variation in selection of peak start and end point. The advantage of the proposed method was that it uses peak shape information of three individual peaks to determine points for integration. That could be the reason for improved RSD percentages compared against the other methods. The root cause for the greater improvement with high RSDs with the current method could be investigated in the future by observing the peak shapes more closely.

References

References

- [1] R. N. Xu, L. Fan, M. J. Rieser, and T. A. El-Shourbagy. Recent advances in high-throughput quantitative bioanalysis by lc-ms/ms. *Journal of pharmaceutical and biomedical analysis*, 44(2):342–355, 2007. URL www.scopus.com. Cited By :417.
- [2] S. Castillo, P. Gopalacharyulu, L. Yetukuri, and M. Orešič. Algorithms and tools for the preprocessing of lc-ms metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, 108(1):23–32, 2011. URL www.scopus.com. Cited By :117.
- [3] M. Katajamaa and M. Orešič. Processing methods for differential analysis of lc/ms profile data. *BMC Bioinformatics*, 6, 2005. URL www.scopus.com. Cited By :308.
- [4] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak. Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006. URL www.scopus.com. Cited By :2846.
- [5] D. Broadhurst, R. Goodacre, S. N. Reinke, J. Kuligowski, I. D. Wilson, M. R. Lewis, and W. B. Dunn. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, 14(6), 2018. URL www.scopus.com. Cited By :264.
- [6] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC bioinformatics*, 10(1):1–13, 2009.
- [7] Robert E Ardrey. *Liquid chromatography-mass spectrometry: an introduction*, volume 2. John Wiley & Sons, 2003.
- [8] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka. Recent advances and prospects of computational methods for metabolite identification: A review with emphasis on machine learning approaches. *Briefings in Bioinformatics*, 20(6):2028–2043, 2019. URL www.scopus.com. Cited By :23.
- [9] J. W. Allwood and R. Goodacre. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis*, 21(1):33–47, 2010. URL www.scopus.com. Cited By :171.
- [10] James J Pitt. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist Reviews*, 30(1):19, 2009.

- [11] Helen Stahnke, Stefan Kittlaus, Günther Kempe, Christlieb Hemmerling, and Lutz Alder. The influence of electrospray ion source design on matrix effects. *Journal of Mass spectrometry*, 47(7):875–884, 2012.
- [12] N. B. Cech and C. G. Enke. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass spectrometry reviews*, 20(6):362–387, 2001. URL www.scopus.com. Cited By :1026.
- [13] K. Jindal, M. Narayanam, and S. Singh. A systematic strategy for the identification and determination of pharmaceuticals in environment using advanced lc-ms tools: Application to ground water samples. *Journal of pharmaceutical and biomedical analysis*, 108:86–96, 2015. URL www.scopus.com. Cited By :27.
- [14] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher. Openms: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741–748, 2016. URL www.scopus.com. Cited By :285.
- [15] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič. Mzmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11, 2010. URL www.scopus.com. Cited By :1880.
- [16] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. Vandergheynst, O. Fiehn, and M. Arita. Ms-dial: Data-independent ms/ms deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6):523–526, 2015. URL www.scopus.com. Cited By :929.
- [17] Eugene Melamud, Livia Vastag, and Joshua D Rabinowitz. Metabolomic analysis and visualization engine for lc- ms data. *Analytical chemistry*, 82(23):9818–9826, 2010.
- [18] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current genomics*, 10(6):388, 2009.
- [19] A. H. P. America and J. H. G. Cordewener. Comparative lc-ms: A landscape of peaks and valleys. *Proteomics*, 8(4):731–749, 2008. URL www.scopus.com. Cited By :161.
- [20] David C Hoaglin, Frederick Mosteller, and John W Tukey. *Understanding robust and exploratory data analysis*. Number Sirsi) i9780471384915. 2000.

- [21] Eva Lange, Clemens Gröpl, Knut Reinert, Oliver Kohlbacher, and Andreas Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. In *Biocomputing 2006*, pages 243–254. World Scientific, 2006.
- [22] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006. URL www.scopus.com. Cited By :472.
- [23] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. Highly sensitive feature detection for high resolution lc/ms. *BMC bioinformatics*, 9(1):1–16, 2008.
- [24] P. G. Stevenson, F. Gritti, and G. Guiochon. Automated methods for the location of the boundaries of chromatographic peaks. *Journal of Chromatography A*, 1218(45):8255–8263, 2011. URL www.scopus.com. Cited By :43.
- [25] Valerio B Di Marco and G Giorgio Bombi. Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*, 931(1-2):1–30, 2001.
- [26] P. G. Stevenson, H. Gao, F. Gritti, and G. Guiochon. Removing the ambiguity of data processing methods: Optimizing the location of peak boundaries for accurate moment calculations. *Journal of Separation Science*, 36(2):279–287, 2013. URL www.scopus.com. Cited By :28.
- [27] P. Nikitas, A. Pappa-Louisi, and A. Papageorgiou. On the equations describing chromatographic peaks and the problem of the deconvolution of overlapped peaks. *Journal of Chromatography A*, 912(1):13–29, 2001. URL www.scopus.com. Cited By :88.
- [28] Ł. Komsta. Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia*, 73(7-8):721–731, 2011. URL www.scopus.com. Cited By :38.
- [29] Gerald A Pearson. A general baseline-recognition and baseline-flattening algorithm. *Journal of Magnetic Resonance (1969)*, 27(2):265–272, 1977.
- [30] Alvin W Moore and James W Jorgenson. Median filtering for removal of low-frequency background drift. *Analytical chemistry*, 65(2):188–191, 1993.
- [31] Andreas F Ruckstuhl, Matthew P Jacobson, Robert W Field, and James A Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, 2001.

- [32] Feng Gan, Guihua Ruan, and Jinyuan Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):59–65, 2006.
- [33] Michał Daszykowski and Beata Walczak. Use and abuse of chemometrics in chromatography. *TrAC Trends in Analytical Chemistry*, 25(11):1081–1096, 2006.
- [34] Miroslav Morháč, Ján Kliman, Vladislav Matoušek, Martin Veselský, and Ivan Turzo. Background elimination methods for multidimensional coincidence γ -ray spectra. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 401(1):113–132, 1997.
- [35] X. Ning, I. W. Selesnick, and L. Duval. Chromatogram baseline estimation and denoising using sparsity (beads). *Chemometrics and Intelligent Laboratory Systems*, 139:156–167, 2014. URL www.scopus.com. Cited By :102.
- [36] Henri P Gavin. The levenberg-marquardt algorithm for nonlinear least squares curve-fitting problems. *Department of Civil and Environmental Engineering, Duke University*, 19, 2019.