TOR VERGATA
UNIVERSITY OF ROME

UNIVERSITY OF ROME "TOR VERGATA"

PH.D. PROGRAM IN INDUSTRIAL ENGINEERING

CYCLE XXXII - A.A. 2019/2020

**Reliability of Data Acquisition
and Processing Approaches for
Chromatography Detection Systems
in Pharmacy**

PH.D. THESIS

submitted by David Thomas Marehn, M.Sc.

in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Industrial Engineering

Tutor: Prof. Dott. Ric. Roberto Pizzoferrato

Co-Tutors: Prof. Dr. rer. nat. Heike Pospisil
Graduate Chemist Detlef Wilhelm

Coordinator: Prof. Dott. Ric. Marco Marinelli

Berlin, January 15th, 2020

Tutor

Prof. Dott. Ric. Roberto Pizzoferrato
Department of Industrial Engineering
Università degli Studi di Roma “Tor Vergata”
Via Orazio Raimondo 18
00173 Rome, Italy

Co-Tutors

Prof. Dr. rer. nat. Heike Pospisil
High Performance Computing in Life Sciences
Technical University of Applied Sciences Wildau
Hochschulring 1
15745 Wildau, Germany

Graduate Chemist Detlef Wilhelm
AnaTox GmbH & Co. KG
Reinheimer Straße 18b
15517 Fuerstenwalde, Germany

Coordinator

Prof. Dott. Ric. Marco Marinelli
Università degli Studi di Roma “Tor Vergata”
Via Orazio Raimondo 18
00173 Rome, Italy

For my close friend and personal hero, Philipp Ecke

Acknowledgment

First of all, I really would like to thank my supervisors Prof. Dott. Ric. Roberto Pizzoferato, Prof. Dr. Heike Pospisil and Detlef Wilhelm for any spent minute in this presented thesis. I thank them for each offered inspiration, discussion, advice and cooperation. Without their effort and believe in my abilities I would not be able to present this work.

Furthermore, I would like to express my gratitude to Prof. Dott. Ric. Roberto Montanari, Prof. Dott. Ric. Mauro Casalboni and Prof. Dr. Maria Richetta. I thank them for their feedback and proposals over of last three years.

Notably, I would also like to thank the whole working group of Prof. Dr. Pospisil, High Performance Computing in Life Sciences, for all the help and suggestions to improve my presentations and publications.

Besides, I would like to express my deepest thanks to my family for their organizational and emotional support. For this and everything else, I thank them from the bottom of my heart.

Statement of Authorship

I, David Thomas Marehn, herewith affirm that I wrote the present dissertation with the title

Reliability of Data Acquisition and Processing Approaches for Chromatography Detection Systems in Pharmacy

on my own and did not use any unnamed sources or aid. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Berlin, January 15th, 2020

.....
(Signature of Ph.D. student)

Abstract

Nowadays pharmaceutical analysis and industry could not be imagined without using chromatographic methods like the High Performance Liquid Chromatography (HPLC) and Gas Chromatography (GC). Therefore, the field of chromatography is already firmly anchored in the three important, regional pharmacopeias: EP, USP and JP. Deficits of specifications within these pharmacopoeias relating to the detector parameters **sampling rate** and **signal filtration** are the major motivation of this thesis. Furthermore investigations on the data acquisition and processing within the detectors and software products dealing with collected data were done. Several concepts like double entry method, smoothing optimization and signal filtration based on persistence have been developed and are used as tools to examine the data integrity of commercial Chromatography Data System (CDS), improve the signal-to-noise ratio of small peaks to lower the LOD and LOQ by efficient denoising, and determine the suitable filter parameters and sampling rates an user can apply on his system to accelerate the method development and validation.

All developments have been tested on simulated and real chromatograms and have shown that they are suitable for their specific purposes. In the end optimizations in the concepts still exist but some new aspects in the long-term investigated field of chromatography have been discovered.

Contents

1	Introduction	1
1.1	International Pharmacopoeias	1
1.1.1	Harmonization of USP, EP and JP	2
1.1.2	Impurities	3
1.2	Liquid Chromatography (LC) and Gas Chromatography (GC)	3
1.2.1	Chromatography in Pharmacopoeia	3
1.2.2	System Suitability for Chromatography Analysis	4
1.3	Sampling Rate and Signal Filtration	6
1.3.1	Influences on the Chromatographical Performance	7
1.4	Motivation - An Optimization Problem	7
2	Data Integrity	9
2.1	Establishment of Double Entry Method during HPLC Analyzes	9
2.2	Application of the Double Entry Method to Preclude Unwanted Manipulations	11
2.2.1	Analytical Instruments	12
2.2.2	Controller	13
2.2.3	Chemicals	13
2.2.4	Software and Chromatographic Conditions	14
2.2.5	Experimental Setup	14
2.3	Preclusion of Data Manipulation by the <i>Data Collector</i> Communication Library	15
2.4	Peak Detection and Evaluation	16
2.4.1	Peak Detection via <i>Apex Track</i> Principle	16
2.4.2	Peak Evaluation	18
2.5	Results	20
2.5.1	Review of Sampling Rate and Signal Filtration	20
2.5.2	Comparison of Simultaneously Acquired Data by the Double Entry Method	23
2.5.3	Preclusion of Data Manipulation by Licop Library via Data Sniffer . .	27
2.5.4	In-Depth Analysis of Data Acquisition by Chromeleon via Data Sniffer	28
2.6	Discussion	30
2.6.1	Sampling Rate and Signal Filtration	30
2.6.2	Integrity and Comparability of Acquired Data by CDS	30
2.7	Summary	32
3	Multi-Objective Optimized Signal Smoothing	33
3.1	Concept of Recommendation	33
3.2	Functionality 1 - Reduction of the Sampling Rate	35

3.3	Functionality 2 - Estimation of Hardware Based and Performing of Firmware Based Signal Filtration	35
3.3.1	Hardware Based Filter	35
3.3.2	Firmware Based Window Filter	36
3.4	Functionality 3 - Multi Objective Optimization	38
3.4.1	Pareto Optimization	39
3.4.2	Parameter Range of Interest	40
3.5	Functionality 4 - Active Substance Analysis versus Purification Analysis . . .	42
3.6	Recommended Parameters - How to realize them?	43
3.7	Simulation of Chromatograms	44
3.8	SmooSi Maker - <u>Smooth Signal Maker</u>	45
3.8.1	Optimization of Chromatograms via SmooSi Maker	45
3.8.2	Suitability of the Response Time Parameter	46
3.9	Results	47
3.9.1	Optimization of Signal Smoothing	47
3.9.2	Response Time Recommendation	49
3.10	Discussion	51
3.11	Summary	53
4	Persistence Filter	54
4.1	Step 1: Filtration of not Relevant Data	55
4.2	Step 2: Reconstruction of the Gaps using Laplace Interpolation	57
4.2.1	Quadratic Programming	58
4.2.2	Iterative Preconditioned Biconjugate Gradient Method (PBCG) . . .	60
4.2.3	Biharmonic and Triharmonic Reconstruction	61
4.2.4	Weighted Inclusion of Original Data into the Reconstruction	63
4.3	Editable Parameters	65
4.4	Results	66
4.4.1	Influence of Data Weight (DW) Value	66
4.4.2	Application of Persistence Filter on Large Peaks	67
4.4.3	Threshold versus Data Weight when Applying Persistence Filter on Small Peaks	68
4.5	Discussion	69
4.6	Summary	71
5	Conclusion and Outlook	72
	Publications	74
	Bibliography	75
	List of Figures	81
	List of Tables	83
	Appendix	84
A.1	Peak Efficiency of Data Acquired by ADC	84
A.2	Data Packages Caught from the Ethernet Connection	85

A.3	In-Depth analysis of Chromeleon®	90
A.4	Chromatogram Simulation Parameters	91

List of Abbreviations

ADC Analog-To-Digital Converter

ASTM American Society for Testing and Materials

CDS Chromatography Data System

CG Conjugate Gradient Method

CFR21 Code of Federal Regulations Title 21

DAD Diode Array Detector

DW Data Weight

EP European Pharmacopoeia

GC Gas Chromatography

HP Hewlett-Packard

HPLC High Performance Liquid Chromatography

ICH International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use

JP Japanese Pharmacopoeia

LC Liquid Chromatography

LOD Limit of Detection

LOQ Limit of Quantification

mAU Milli Absorbance Unit

PBCG Preconditioned Biconjugate Gradient Method

PDG Pharmacopoeial Discussion Group

POF Pareto Optimal Front

RAM Read Access Memory

RSD Relative Standard Deviation

USP United States Pharmacopoeia

WS Window Size

1 Introduction

CHROMATOGRAPHY, a term named by Michail Semjonowitsch Tswett in the year 1903 [1], developed to one of the major separation processes of the present during the last century. Many different types of chromatography have been researched, optimized and established in a wide range of applications. Nowadays, the chromatography is daily used in the forensic, food regulation, drug testing and quality control [2]. Especially the pharmaceutical analysis and industry could not be imagined without chromatographic methods like the *Liquid Chromatography* and *Gas Chromatography*. Therefore both methods are well described and firmly anchored in the three most important, regional pharmacopoeias: United States Pharmacopoeia [3], European Pharmacopoeia [4] and Japanese Pharmacopoeia [5]. These pharmacopoeias and deficits in their monographs and general chapters about the chromatography are the motivation of this presented Ph.D. thesis.

The sections 1.1 to 1.4 describe the content and functionality of the international pharmacopoeias. A closer examination of the general chapters dealing with chromatography leads to the demand for investigations on the detection parameters **sampling rate** and **signal filtration** which are the essential topics of interest of the thesis. In chapter 2, the present manuscript covers the theory and conducted pretests which verify traceability and applicability of any further experiences due to the established equipment and measurement systems that are available to laboratory assistants. Chapter 3 is directly addressed to the parameters **sampling rate** and **signal filtration** and possible improvements of their usage in the regulated environment. For that purpose different types of filters for smoothing and denoising acquired chromatographic signals have been investigated by a new self-written software project called *SmooSi Maker*. The results here inspired to develop a *Persistence Filter* based on knowledges from the fields of geology and image processing in order to denoise the chromatograms on condition that peak information are influenced as little as possible. This topic is handled in chapter 4.

1.1 International Pharmacopoeias

Dealing with drugs requires thorough manufacturing, quality control and medication. That is why different organizations have been founded to collect and write down the best practices for each step in order to make an uniform guideline. Several pharmacopoeias were created depending on the different places of origin. The following three ones belong to the global

players:

United States Pharmacopoeia (USP) The USP is managed by the *United States Pharmacopoeial Convention* and is the most far-reaching pharmacopoeia applied in over 140 countries [6].

European Pharmacopoeia (EP) The EP currently has 38 members and 27 observers in countries especially of Europe and Africa [7]. It is managed by the *European Directorate for the Quality of Medicines & HealthCare*.

Japanese Pharmacopoeia (JP) The JP is managed by the Japanese *Ministry of Health, Labour and Welfare*. It plays an important role and is involved in the harmonization of the different pharmacopoeias (section 1.1.1) even if the JP is applied in Japan only [6].

Each pharmacopoeia contains information about drugs (obtainable with or without a prescription), excipients and supplements covering their active substances, tests for critical quality attributes and applications [6]. All these information are summarized in so called monographs. One monograph deals with one drug, excipient or supplement [8]. In addition, the pharmacopoeias contain general chapters. All the analytical methods and approaches that are required to meet conditions of the monographs are described and specified within these chapters [8]. The chromatography itself also has an own general chapter (section 1.2.1).

1.1.1 Harmonization of USP, EP and JP

All three big organizations have similar intentions when creating and publishing their pharmacopoeias but there are unavoidable deviations due to the regional differences. Because of the present globalization the desire of a harmonization between USP, EP and JP came up. Therefore the *Pharmacopoeial Discussion Group (PDG)* has been founded and the program *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)* was set up in order to initiate and observe the harmonization [9]. A good harmonization would lead into a more efficient global drug supply and avoid redundant testings.

Every monograph or general chapter that will be harmonized has to complete the seven stages of the PDG working procedure [10] whereby the seventh and last stage is the inter-regional acceptance. That means every organization like the *United States Pharmacopoeial Convention* primarily aims to target the stage six before it signs off the corresponding monograph or general chapter [11]. So far, the general chapter dealing with the chromatography passed stage four. That means this chapter is midst of the harmonization process and there are still deviations between the USP, EP and JP (section 1.2.1).

1.1.2 Impurities

One major aspect that is considered in all the monographs of the medical substances is the formation of impurities during the synthetical route of the drug. These impurities are unwanted chemicals which remain with the desired active substance [12]. Their presence disturb the desired pureness of the active substance which is necessary to avoid unexpected up to harmful effects when the drug is medicated by a patient.

For that reason all known and unknown impurities of an active substance in the pharmacopoeias are listed [13]. In order to avoid or minimize impurities, suitable impurity tests using reference standards are defined. Performing these tests allow to make statements about the concentrations of impurities and if they exceed maximal permitted acceptances. Generally, peaks that areas are below a limit of 0.05% related to the area of the principal peak in the specific reference standard are disregarded [3]. All impurities listed in the monographs belong to one specific synthetic route [13]. That means for a new route the generated impurities may not be covered by the monographs.

1.2 Liquid Chromatography (LC) and Gas Chromatography (GC)

The chromatography belongs to the field of analytical chemistry and is a separation technique based on the chemical and physical interaction between the components of interest with a mobile and a stationary phase of the analytical system. Differences in the distribution coefficient of similar molecules are used to separate them. The choice of mobile and stationary phase decides about the kind of chromatography. The GC consists of a solid stationary phase and a gaseous mobile phase. The molecules separated by this method will be mostly evaporated previously. For the LC a liquid mobile phase is used. The stationary phase is still solid. Solved molecules will be separated when passing the stationary phase. There are also other variations like a liquid stationary phase used for the GC and other types of chromatography like the thin-layer or paper chromatography [3].

1.2.1 Chromatography in Pharmacopoeia

Every established type of chromatography is listed and described in the USP, EP and JP containing the following information [3–5]:

- Explanation and specification of the chromatography type. What are the mobile and stationary phases?
- Exact description how to perform the chromatography technique.
- Definition and determination of specific values used for the evaluation of the obtained results. Any required formulas are written down and explained.

- Specification of tests to ensure the suitability of the chromatographic system to the analysis it is designed for.

In the general chapters describing LC and GC there are listed all formulas that are needed to evaluate a so called chromatogram, which will be acquired during an experiment. Such a chromatogram is presented in fig. 1.1. In general a chromatogram is a representation of the behavior between the analysis time and a signal/response produced by the detector unit of the chromatographic system. A chromatogram starts with the injection time t_0 up to a previously defined run time. In LC there is often an air peak in the chromatogram due to the injection. The air is not retained by the system and that is why the time of the peak after injection is called hold-up time t_M which is not shown in the figure. It is the minimal time a sample needs to get from the injector to the detector depending on the flow of the mobile phase and a few other parameters excluding the stationary phase. This also applies to any other components than air that are part of the sample like a solvent which is not retained by the system. An other really important time value is the retention time t_R . This corresponds to the time from injection up to peak apex of a component in the sample. For an unaltered system the retention time would be always the same and can be used for the identification of a component. The previously mentioned term "peak" is a change in the signal depending on a component passing the detector measurement cell. Such a peak has several properties [3] like a height H (between baseline and peak apex) and a width W . Due to an easier measurement the peak width at the half height ($W_{0.5}$ or $W_{1/2}$) is mostly used for calculations. Further terms that are required for the evaluation are peak-to-peak noise (h), peak width at 5% height ($W_{0.05}$) and the distance from the peak maximum to the leading edge of the peak (f).

There are several differences in the pharmacopoeias dealing with these information due to the fact that the chapter of chromatography is still in process of harmonization [11]. One example is the formula calculating the number of theoretical plates (peak efficiency) using the retention time and peak width of a component in the chromatogram. The EP and JP specify a formula using the peak width at the half height (eq. (2.4), page 19) [4, 5] but the USP additionally lists and recommends an other one using the peak width at the baseline [3]. A similar behavior applies for the calculation of peak resolution (eq. (2.7), page 20).

The USP has the most members and that is why these deviations will be excluded by concentrating firstly on this pharmacopoeia when having a closer examination on the part this thesis is interested on: the system suitability.

1.2.2 System Suitability for Chromatography Analysis

The system suitability section of the pharmacopoeias describes tests as an internal part of real chromatography analysis. These tests are used to guarantee an adequate performance

Figure 1.1: Several terms for description and evaluation of a chromatogram

The shown chromatogram summarizes several terms based equations written down in the general chapters of USP that are used to evaluate the suitability of the chromatographic system [3]. The presented terms are the retention times t_{R1} and t_{R2} , the peak height H , the peak width at half height $W_{1/2}$ (also $W_{0.5}$), the peak-to-peak noise h , the peak width at 5% height $W_{0.05}$ and the distance from the peak maximum to the leading edge of the peak f .

of the given system for the intended use [3–5]. One example is the test parameter ”signal-to-noise ratio”. Here the peak height ($\hat{=}$ signal) and the background noise in a well defined range of baseline signal ($\hat{=}$ noise) are used to compute the signal-to-noise ratio which influences the precision of quantification [4] (eq. (2.6), page 20).

Additionally to the tests and their parameters, the system suitability section deals with the possibility to adjust chromatographic conditions of a monograph. These adjustments allow editing parameters like the column temperature or flow rate of the mobile phase which are usually given in the corresponding monograph in order to fulfill or improve the performance. These adjustments may not be done arbitrarily. In the USP and EP every parameter is listed with an allowed range of changes (mostly given in per cent) [3, 4]. Having a closer examination on these parameters, it can be seen that every module in a LC and GC device is affected: the control unit of the flow rate, the injection unit, the column compartment and the detection unit. But two parameters that can be modified by the user and have proven influences on the system performance are missing here. These parameters are the **sampling rate** and **signal filtration** of the detector unit.

1.3 Sampling Rate and Signal Filtration

When a chromatographic system will be used and established for a single experiment or experiment sequences in order to qualify and/or quantify one up to several compounds of interest in an unknown sample, the user has to make a choice which parameters would be the best for his intentions. One way is to refer on experiences of previous experiments. In the regulated environment the user has to refer to the well defined specifications in the pharmacopoeias when he analyses a component listed in the monographs (section 1.1). Otherwise the user has to design experiments by himself. This will end up to a validation and verification procedure where the user has to clarify the intended usage of his analysis and find the suitable parameters afterwards.

Beside the obvious reliance on used hardware the user has to define a suitable method. The method of a chromatographic experiment contains parameters like the flow of the mobile phase, injection volume of the sample, temperature of the mobile phase when passing the stationary phase and detection parameter. The most apparent detection parameter is the wavelength when an optical technique is used. For other techniques like refractive index (LC) or thermal conduction (GC) detection the temperature in the measuring chamber is the major parameter. Sampling rate and filtration of the signal are detection parameters, too. But their influences are either underestimated or completely unknown [14, 15].

Regardless of the detection type every detector produces an analog, electrical signal such as changes of voltage or current [16, p. 19]. Usually an internally installed analog-to-digital converter (ADC) digitizes the analog, continuous signal into a discrete signal and provides it to an external controller (e.g. computer) by a suitable interface. The **sampling rate** is a parameter of the ADC and specifies how many discrete data points per second will be sampled and provided. The unit of the sampling rate is given in Hz (1/s). The reciprocal of the sampling rate is the sampling period. Presently there are two approaches how an ADC operates. Either the analog signal will be exactly sampled by the given frequency or the highest available frequency will be used and a specific amount of data points will be averaged to one point. The second method is called data bunching [16, p. 21].

Changes in the signal during an experimental analysis do not just depend on the sample passing the detector. That is because every analog signal is overlaid by an (primarily electronic) noise [16, p. 21]. Noise is considered as a disturbance due to the influence on the accuracy and precision of a chromatographic analysis [17]. Therefore there are accepted values for the noise depending signal-to-noise ratio in pharmaceutical and bio-analytical analysis. The minimal signal-to-noise ratio S/N depends on the desired precision given as Relative Standard Deviation (RSD) and is calculated by:

$$\%RSD = \frac{50}{S/N} \quad (1.1)$$

That means for example a desired precision of 2% requires at least a signal-to-noise ratio of 25. Furthermore, the signal-to-noise ratio is related to the terms Limit of Detection (LOD)¹ and Limit of Quantification (LOQ)². They are preferably defined as a multiple of the signal-to-noise ratio S/N [17]. A rule of thumb defines the LOD as $3 \cdot S/N$ and the LOQ as $10 \cdot S/N$ [18]. Depending on the required precision the factors differ and may increase significantly. If low LOD and LOQ values are desired it is necessary to reduce the baseline noise and/or to increase the signal of the chromatographic system [17].

In order to improve the signal-to-noise ratio a filtration is performed by digital filters during the data acquisition [14] or by filter algorithms during the data processing afterwards [20]. Neither the sampling rate nor the signal filtration should be set arbitrarily (section 1.3.1).

1.3.1 Influences on the Chromatographical Performance

When creating a method for a chromatographical analysis in most cases the user has no idea which sampling rate and filtration parameter should be set. The problem is that the sampling and filtration procedures of the instruments generally are not defined very well by the manufacturers [14] and so the user has to try out.

Several investigations have been done to find out how real chromatograms are affected by the sampling rate and signal filtration. It has been found out that a lower sampling rate decreases the standard deviation of the noise and for a system with a high column performance it decreases the peak efficiency given by the number of theoretical plates [14]. A very low sampling rate can lead to unreproducible chromatograms with changes in peak height and shape for each run due to the sampling theorem [21]. Investigations on hardware digital filters of detectors used for the LC have shown that a high filter parameter given as time constant or response time (depending on hardware manufacturer) will lead to an improved signal-to-noise ratio (as expected) but also to distorted peaks [15, 20] up to a tailing phenomenon of the peaks which not depends on the residence time of the components in HPLC separation column [14].

1.4 Motivation - An Optimization Problem

Based on the previous sections the question comes up if it is possible to find optimal sampling rate and signal filtration parameter settings and advice them to an user or a regulatory

¹The point at which analysis of the desired component is just feasible [18, 19]

²The point at which quantitative results of the desired component can be reported with a high degree of confidence [18, 19]

authority. The term optimal is not connected here to storage size which could limit the sampling rate but with other values that are related to the suitability terms of a chromatogram (fig. 1.1). A suitable improvement procedure would have several advantages to further possible situations.

1. Within the harmonization process mentioned in section 1.1.1 it could happen that changes to the monographs are necessary and providing suitable detector parameters could be included into the pharmacopoeias.
2. When a new monograph will be integrated into the pharmacopoeias the detector parameters could be integrated here, too.
3. Beside the pharmacopoeias optimization algorithms for the sampling rate and signal filtration could be used during any kind of method development and validation. That way developments of a new or reconfigurations of existing methods could be accelerated relating to the detector setups.

Additionally in the field of pharmacy the detection of impurities may be improved due to the influence of denoising on the Limit of Detection (LOD) and Limit of Quantification (LOQ). A good LOD for example has positive effects on the detection of impurities which generates small peaks and may not differ to the baseline noise in a significant way.

2 Data Integrity

In chapter 1 it has been clarified why investigations on the HPLC detector unit parameters sampling rate and signal filtration have to be done. But before any recommendations of these parameters can be provided to any user in regulated environment as such in the field of pharmacy, some pending questions relating to the established Chromatography Data System (CDS), which control the hardware and acquire the provided raw data, have to be answered first.

Questions

1. What is the usual practice for the acquisition of raw data when using chromatographical separation techniques like High Performance Liquid Chromatography (HPLC) and Gas Chromatography (GC)?
2. Is the raw data pre-processed in any way before storing it on the file system or a database? In other words, is it possible to compare the exported raw data of two different CDS and the exports will be completely equal?
3. Is it possible to check the incoming data that will be handled by a usually used commercial CDS in order to evaluate its data integrity?
4. Can the influences of the sampling rate/signal filtration on chromatograms be reproduced?

This chapter deals with the given questions and explains how a new kind parallel data acquisition could be established in order to collect data of a one HPLC system during a single analytical run twice.

2.1 Establishment of Double Entry Method during HPLC Analyzes

The software supplied by a High Performance Liquid Chromatography (HPLC) vendor is called Chromatography Data System (CDS). Beside the main functionalities like instrument control by well developed drivers, peak detection and integration of the acquired chromatograms, the data integrity is an important feature the software has to fulfill in regulated

environments. The term data integrity is defined as the maintenance of accuracy and consistency of acquired raw data over its entire life-cycle [22] and software packages that are designed and developed in the environment of food and drugs are requested to consider this aspect by implementing the guidelines of Code of Federal Regulations Title 21 (CFR21) [23]. This fact also includes that the data acquired from the instrument and the data stored on the file system have to be the same in order to provide them to any user if requested. This behavior is the main motivation in this chapter and as mentioned in question 3 above there is an approach required for a review.

One possible approach that has been developed at this place is inspired by the double entry method used for manual data entries where a data verification is necessary. The data verification is a strategy that ensures that incoming data matches the original one [24]. For manual entries there are two usually used methods dealing with the data verification. One user could either proofread the entered data using the original data as a reference or two separate persons enter the data twice and compare them afterwards. The latter is called double entry method and its principle can be used for automatically acquired data, too. On the other side, the proofread method is not suitable here because usually the data provided by the HPLC instrument will not be stored in a buffer within the instrument. That means the data are gone after the analytical run and are not available as a reference that way. The question now is how to realize such a double entry method for a HPLC analysis.

Figure 2.1: Concept of the simultaneous data acquisition

At the same time, a self-written software with a limited range of functionalities will be started on the computer together with the commercial CDS. The CDS will configure all modules of the HPLC system relating to the configuration of the experiment, start the analysis and acquire the generated and provided data from the detector. In parallel, the self-written tool called *Data Collector* collects data from the detector using the same interface as like as the commercial CDS.

In practice laboratory assistants use only one CDS for an analytical experiment with a chromatographic system. One possible idea could be to run analysis with at least two software packages acquiring data from the detector unit simultaneously. But comparing the acquired data of two established, commercial CDSs such as for example *OpenLab ChemStation*® developed and published by *Agilent Technologies* [25] and *Chromeleon*® developed

and published by *Thermo Fischer* [26] directly is impossible. That is because both CDSs will take the full control of the HPLC system and lock it. That way no other controller can modify the experimental method or general system setup and external manipulations will be avoided. But the lock does not affect the informational output of the system like status monitoring or data acquisition¹. Based on the fact that data acquisition is still possible after locking the HPLC system a parallel data acquisition would be able using a controller software with a lower range of functionalities. The concept of simultaneous data collection by such a data collector tool is shown in fig. 2.1. The required software tool has been written by myself and is simply called *Data Collector* relating to its task.

The revision 1.0 of the *Data Collector* is based on the *Microsoft .NET 4.0* framework and written in C#. For the communication with two available HPLC systems from the 1260 and 1200 series manufactured by *Agilent Technologies* the public available dynamic library *Licop* has been integrated into the *Data Collector*. The *Licop* library is part of the official "LICOP components for software development" published by *Agilent Technologies* [27]. The only precondition for the usage of the *Data Collector* is either one network card installed in the HPLC system that allows two instances for communication or two available network cards installed in two different modules of the system. The usage of the *Data Collector* is divided into five steps:

- 1: Import functionalities of the official LICOP library.
- 2: Connect to HPLC system by a user given IP address or hostname.
- 3: Search for at least one detector in the whole HPLC system.
- 4: Let the user select an output folder for the incoming data.
- 5: Instruct the detector to provide data to the raw data channel of the LICOP library.
- 6: Wait for an external start signal and parse the incoming binary data afterwards until a stop signal occurs.

2.2 Application of the Double Entry Method to Preclude Unwanted Manipulations

In this section the experimental conditions are listed to check the data integrity and to review the influence of the sampling rate and signal filtration. All instruments and materials listed in this section were provided by the company *AnaTox GmbH & Co. KG*.

¹That has been found out by connecting an external panel such as the instant pilot G4208A (*Agilent Technologies*). This panel still gets the actual status and provided data of the device even after locking the system by a connected CDS.

2.2.1 Analytical Instruments

All the experiments were done and repeated on two several HPLC systems. Both systems are manufactured by *Agilent Technologies*. In the further course of this thesis one system will be called 1260 series and the other one 1200 series system corresponding to the used modules. The systems are defined as follows.

1260 series HPLC system

- G4225A Degasser
- G1312B Binary Pump (Revision A.06.54)
- G1367E Wellplate Autosampler (Revision A.06.54)
- G1330B Autosampler Thermostat
- G1316C Column Compartment (Revision A.06.53)
- G4212B Diode Array Detector (DAD) (Revision B.06.72)

1200 series HPLC system

- G1322A Degasser
- G1311A Quaternary Pump (Revision A.06.55)
- G1329A Standard Autosampler (Revision A.06.54)
- G1316A Column Compartment (Revision A.06.50)
- G1315D DAD (Revision B.06.52)

Furthermore, the detector of the HPLC system has been connected to an external analog-to-digital converter (ADC). The manufacturer of this ADC is *Agilent Technologies*, too. Its properties are the following.

35900E Dual Channel Interface

- Conversion of up to two incoming analog signals
- Sampling rate from 0.1 to 100 Hz
- Accepted input voltage from -18 millivolts to 1 volt

2.2.2 Controller

All the experimental sequences have been designed and performed on a *Hewlett-Packard Compaq dc 7800 Small Form Factor* computer (product number: KS467UP#AK6) with an Intel(R) Core 2 Duo CPU E6550 @ 2.33 GHz processor and 4096 MB RAM. The running operation system was Windows 7 Professional SP1 32-bit.

Figure 2.2: Exact experimental setup of the parallel data acquisition

The setup used for the CDS *OpenLab ChemStation*[®] and *Chromeleon*[®] differ due to the available communication with the analog-digital-converter. For both CDSs the data of the HPLC detector provided by the Ethernet connection have been acquired by the CDS itself and the *Data Collector*. In addition data of the analog-digital-converter have been acquired by the CDS *OpenLab ChemStation*[®].

2.2.3 Chemicals

An isocratic test sample supplied by *Agilent Technologies* (product number: 01080-68704) containing four substances has been used for the experiments due to an isocratic and isotherm configuration. The substances are:

- Dimethyl phthalate
- Diethyl phthalate
- Biphenyl
- o-Terphenyl

These components were solved in Methanol. For the separation of the components a mix of 35 vol.% HPLC-grade water and 65 vol. % Acetonitrile was used as mobile phase and a Zorbax xDB-C8 column supplied by *Agilent Technologies* (product number: 922975-906)

was installed as stationary phase according to a reverse phase chromatography configuration. The column had a length of 50 mm, a diameter of 4.6 mm and a pore size of 1.8 μm .

2.2.4 Software and Chromatographic Conditions

The methods of the chromatographic experiment were set up by the CDS *OpenLab ChemStation*[®] (Rev. C.01.07 [27]) developed and published by *Agilent Technologies* or *Chromeleon*[®] (Rev. 6.80 SR15 Build 4656) developed and published by *Thermo Fischer*. The exact conditions have been chosen corresponding to the recommended specifications on the certificate of the isocratic test sample (section 2.2.3) and are listed in table 2.1.

Table 2.1: Chromatographic conditions for simultaneously data acquisition

The chromatographic conditions for both HPLC systems 1260 and 1200 series are primarily equal except the available settings for the sampling rate and digital filtration (given as response time). For the present systems it is not possible to set up the sampling rate or filtration independently.

Parameter	1260 series HPLC system	1200 series HPLC system
Run Time [min]	5	5
Flow [ml/min]	1	1
Eluent Gradient	isocratic	isocratic
Injection Volume [μl]	1	1
Sampler Draw Speed [$\mu\text{l}/\text{min}$]	50	50
Sampler Eject Speed [$\mu\text{l}/\text{min}$]	200	200
Column Temperature [$^{\circ}\text{C}$]	40 $\pm$ 0.5	40 $\pm$ 0.5
Detector Signal [nm]	254	254
Detector Signal Bandwidth [nm]	4	4
Detector Reference [nm]	500	500
Detector Reference Bandwidth [nm]	100	100
Offset [%]	5	5
Sampling Rate [Hz] / Response Time [s]	80 / 0.031	20 / 0.13
	80 / 0.063	20 / 0.25
	40 / 0.13	10 / 0.5
	20 / 0.25	5 / 1
	10 / 0.5	2.5 / 2
	5 / 1	1.25 / 4
	2.5 / 2	0.62 / 8
	1.25 / 4	0.31 / 16
	0.62 / 8	-
	0.31 / 16	-

2.2.5 Experimental Setup

The exact experiment setup consisting of the analytical instruments 1260 and 1200 series HPLC system (section 2.2.1) and the controlling CDSs *OpenLab ChemStation*[®] and

Chromeleon[®] are shown in fig. 2.2. Every available sampling rate and signal filtration mode (table 2.1) has been configured and each kind of configuration has been repeated for ten times.

This ended up into

1260 series: 2 CDS · 10 configurations · 10 repetitions = 200 runs

1200 series: 2 CDS · 8 configurations · 10 repetitions = 160 runs

$$\sum = 360 \text{ runs}$$

Additionally, some runs have been repeated in order to cover all available sampling rates of the analog-digital-converter (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50 and 100 Hz).

2.3 Preclusion of Data Manipulation by the Data Collector Communication Library

The own written *Data Collector* uses the official *Licop* library [27] as a driver for the hardware communication. That means the communication to the HPLC system is not completely understood at this point and a possible manipulation on the acquired data can not be precluded per se because the data always pass the *Licop* library first.

Figure 2.3: Bypass the Licop library

A simple experiment setup using the *Data Collector* is shown. The *Data Collector* bases on the *Licop* library [27] in order to communicate with the HPLC system correctly. Therefore the acquired data from the detector will pass the library before the tool can parse the data itself. A second tool called *WireShark* [28] has been used to observe the Ethernet connection and catch the data packages before they reach the library.

In order to investigate the behavior of the library a simple setup has been used (fig. 2.3). The test system for the single run was the 1200 series HPLC. Instead of a separation column a restriction capillary has been installed and the eluent was HPLC-grade water. A solution of 25 µg/ml Caffeine solved in water has been used for the injection. All other chromatographic conditions of the method configured by an *Agilent Technologies* G4208A instant pilot are listed in table 2.2.

Just before the experimental run an additional tool has been started: *WireShark* (Rev. 2.0.5) [28]. This tool allows to observe data packages transmitted in the local area network. That way it was possible to bypass the data packages from the system to the controller (here the computer) directly. See the red arrow labeled by "Ethernet" in fig. 2.3. This data did not pass the *Licop* library.

Table 2.2: Single run conditions for Licop checkup

The chromatographic conditions for the single run using the 1200 series HPLC system and a 25 µg/ml sample of Caffeine for the injection.

Parameter	1200 series HPLC system
Run Time [min]	1
Flow [ml/min]	1
Eluent Gradient	isocratic
Injection Volume [µl]	10
Sampler Draw Speed [µl/min]	50
Sampler Eject Speed [µl/min]	200
Column Temperature [°C]	25 ±0.5
Detector Signal [nm]	273
Detector Signal Bandwidth [nm]	4
Detector Reference [nm]	500
Detector Reference Bandwidth [nm]	100
Offset [%]	5
Sampling Rate [Hz]	2.5
Response Time [s]	2

2.4 Peak Detection and Evaluation

This section summarizes how all resulting chromatograms in this thesis have been evaluated. The evaluation is divided into the peak detection and the following calculation of the peak describing terms shown in fig. 1.1.

2.4.1 Peak Detection via Apex Track Principle

Several peak detection algorithms have been established by the HPLC and GC vendors during the development and improvement of their systems in order to provide an exact and

automatic identification of the peaks as possible. For any peak detection within this thesis it has been decided to use the *Apex Track* algorithm developed by *Waters Corporation* [29] and implemented into their *CDS Empower*[®] [30]. This approach has been chosen because its algorithm requires only a small number of parameters to regulate it.

The principle of the *Apex Track* is explained in short in fig. 2.4. First of all it is necessary to find the apex of the peak, the highest point. This can be done for example by locating a local minimum within the second derivative of the chromatogram. Due to the fact that the second derivative is much more affected by baseline noise than the original chromatogram [20] it has been decided to search for local maxima within the chromatogram using the watershed transformation². When a peak apex has been found the inflection points will be determined. Therefore the first derivative of the chromatogram is used in order to find the local maximum (inflection point of leading edge) and the local minimum (inflection point of trailing edge) around the apex (fig. 2.4, 1).

Figure 2.4: Peak Detection via *Apex Track*

The peak apex and its inflection points will be determined (1). Starting from the inflections points an initial baseline (2) will move to the bottom (3) until the angles between the baseline and the tangents of the data points at the end of the line fall below a threshold (4). Then the baseline forms the liftoff and touchdown time points of the peak.

The images 2, 3 and 4 in fig. 2.4 explain the second step of the *Apex Track* principle: The baseline expansion. An initial baseline is set between the inflection points (fig. 2.4, 2). Then

²The watershed transformation will be explained in detail by fig. 4.2 on page 56.

the angles between the baseline and the tangents of the points will be determined and used as reference for the following expansion. That means the baseline moves from the inflection points to the bottom and for every move the changed angles will be computed once more and compared with the initial angles (fig. 2.4, 3). If the angles fall below a threshold – which depends on the initial baseline – then the baseline expansion stops and the liftoff and touchdown time points have been determined (fig. 2.4, 4). The borders of the peak have been detected now.

2.4.2 Peak Evaluation

At this place all the terms are listed and described which are necessary to evaluate any chromatogram as required in this thesis. Most of the calculations and determination concepts are taken from the pharmacopoeias [3–5] in order to have a valid fundament all evaluations are build on. The equations themselves are listed here in order to reference them anywhere in the thesis when they are used.

Peak-to-Peak Noise

The baseline noise of a given chromatogram can be quantified by different ways [31] but the so called peak-to-peak noise N_{PP} is the chosen one in the pharmacopoeias. In the desired chromatogram the noise will be determined in a time range that is far away from any peak in order to get the baseline noise that is not influenced. Otherwise a reference chromatogram without any injection and no resulting peak could be used. In the defined area a regression line has to be computed and the peak-to-peak noise is calculated then by the following equation [3].

$$(N_{PP}) = s_{max} - s_{min} \quad (2.1)$$

with

N_{PP} = Peak-to-Peak Noise

s_{max} = Signal value with largest positive distance to the regression line

s_{min} = Signal value with largest negative distance to the regression line

Peak Width at Half Height

Peak width at half height $W_{0.5}$ or $W_{1/2}$ defines the time span between the leading and trailing edges of the peak where the signal crosses the half height points [32]. It is possible to measure the signals at the half height by searching for the data point with the minimal distance to it [14, Supplements], but in this thesis a simple linear interpolation is used in order to be more accurate because during the investigations the sampling rate has been decreased and

the deviation of the nearest data point could be $\frac{1}{2} \cdot \text{sampling period}$ at worst. The estimation of the data point at a specific peak height is summed up in eq. (2.2). For each the leading and the trailing edge of the peak the next data point above ($> H\%$) and below ($< H\%$) will be searched. Then a line will be formed between these points and finally the resulting equation will be rearranged in order to calculate the time for signal at the specific height.

$$t_{H\%} = \frac{s_{H\%} - n}{m} \quad (2.2)$$

with

$$m = \text{scope} = \frac{s_{(<H\%)} - s_{(>H\%)}}{t_{(<H\%)} - t_{(>H\%)}}$$

$$n = \text{intercept}$$

Setting $H\%$ to 50% allows the measurement of the time value of data points at half height and calculate the peak width at half height $W_{0.5}$ by difference [32].

$$W_{0.5} = t_{50\%, \text{trailing edge}} - t_{50\%, \text{leading edge}} \quad (2.3)$$

Column Efficiency

The column efficiency N is a term that makes an assertion about the interaction between the component in the eluent and the used separation column. It is indicated as the number of theoretical plates and it can be calculated after generating a chromatogram by peak width at half height $W_{0.5}$ and retention time t_R [3].

$$N = 5.54 \left(\frac{t_R}{W_{0.5}} \right) \quad (2.4)$$

Symmetry Factor alias Tailing Factor

The symmetry factor A_s is the relation between the peak width at 5% peak height and double of the distance from the peak maximum to the leading edge f [3]. These parameters are visualized in fig. 1.1 (page 5). A symmetry factor of 1 corresponds to an ideal Gaussian peak. Lower and greater values than 1 implicate fronting and tailing phenomena, respectively.

$$A_s = \frac{W_{0.05}}{2f} \quad (2.5)$$

Signal-to-Noise Ratio

The signal-to-noise ratio S/N is a relative term between the peak height H of the peak related to the baseline and half of the peak-to-peak noise h [3]. The term h corresponds to the peak-to-peak noise determination as mentioned in eq. (2.1). As explained in section 1.3 the S/N influences the LOD and LOQ of the chromatographic system.

$$S/N = \frac{H}{\frac{1}{2}h} = \frac{2H}{h} = \frac{2H}{N_{PP}} \quad (2.6)$$

Resolution

The resolution R_s is a value describing the separation between to peaks. The calculation considers the peak weight at half height ($W_{0.5}$) and the retention times (t_R) of peak 1 and 2 [3]. As a rule of the thumb a resolution of at least $R_s = 1.5$ corresponds to a suitable separation [33].

$$R_s = 1.18 \cdot \frac{t_{R2} - t_{R1}}{W_{0.5,1} + W_{0.5,2}} \quad (2.7)$$

2.5 Results

With the implemented *Data Collector* several comparisons have been performed which will be presented in the following section. Parts of the results presented in the sections 2.5.2 to 2.5.4 have been published [34, 35].

2.5.1 Review of Sampling Rate and Signal Filtration

All analytical runs have been done by injecting the isocratic test sample of *Agilent Technologies*. Based on the chosen column and eluent (section 2.2.3) the four components of the isocratic test sample have been separated by performing a reverse phase liquid chromatography. That means the components have been hold back by their nonpolar interaction with the stationary phase when passing the column. The differences of the component polarities led to the specific chromatograms as seen in fig. 2.5. Both chromatograms have been produced by the 1260 series HPLC system (section 2.2.1) and acquired by the self-written *Data Collector*. The experimental conditions between the shown chromatograms are the same except the used sampling rate and digital filtration parameter of the HPLC detector. The first chromatogram (fig. 2.5, above) presents the acquired data using a sampling rate of 2.5 Hz, which corresponds to a sampling period of 400 ms, and a response time of 2 seconds

for the digital filter³. For the second chromatogram (fig. 2.5, below) an 8-fold increased sampling rate has been used whereby the response time was 8-fold decreased. For the given instrument it was not possible to set up the sampling rate and filtration independently.

Figure 2.5: Influences of coupled sampling rate and filtration settings on the chromatogram

Both chromatograms acquired by the *Data Collector* show the separation of the components solved in the isocratic test sample. With increasing retention time the components are Dimethyl phthalate, Diethyl phthalate, Biphenyl and o-Terphenyl. The peak efficiencies (black numbers) decrease when a lower sampling rate and greater response time (2.5 Hz and 2s) are used whereby the retention times (red numbers) are nearly the same. On the other hand the signal-to-noise values S/N_1 to S/N_4 are greater even though the peak heights are reduced.

For each experimental condition 10 repetitions were performed. The average results of the evaluated peaks are directly shown in the fig. 2.5. The black number next to each peak is the column efficiency given as the number of theoretical plates N calculated by eq. (2.4). The retention times are also shown in the fig. 2.5 as the red numbers. Additionally small chromatograms are placed on the right side containing the data of the last 10 seconds. The averaged signal-to-noise ratio values S/N_1 to S/N_4 are computed by the peak heights H_1

³ *Agilent Technologies* uses a moving average filter with Gaussian kernel implemented into their motherboard's firmware [14]

to H_4 and the peak-to-peak noise N_{PP} using eq. (2.6). The peak-to-peak noise has been calculated for the time range from 200 to 300 s.

Whereas the retention times are changing marginally the difference between the peak efficiencies are significant. Relating to eq. (2.4) a configuration with a lower sampling rate and greater response time seems to induce a peak broadening. On the other hand the signal-to-noise ratios increase for the lower sampling rate configuration even though the peak heights are reduced.

The same comparison between two sampling rates has been done for the ADC *35900 Dual Channel Interface*. The result is visible in fig. A.1 in the appendix on page 84. The ADC do not perform any signal filtration, only the sampling rate will be reduced. Here the peak efficiency decreases only a little bit after a 10-fold reduction of the sampling rate. But even without any signal filtration the baseline noise has been reduced for 10 Hz sampling rate.

Figure 2.6: Overlaid plots of *OpenLab ChemStation®* and *Data Collector* for 1260 HPLC series system

For the raw data plots acquired with a sampling rate of 20 Hz (above) no deviations are visible. In the scatter plot (below) of the third peak in the time range from 80 to 120 seconds an increasing value of the signal acquired by *OpenLab ChemStation®* seems to increase the range of difference. Three sections are visible: 0-1 mAU, 1-10 mAU and >10 mAU.

2.5.2 Comparison of Simultaneously Acquired Data by the Double Entry Method

This section deals with the comparison of the simultaneously acquired data when using the self implemented *Data Collector* and the official, commercial Chromatography Data Systems (CDSs). For each HPLC system and CDS the plots of the pure raw data when using a sampling rate of 2.5 Hz and a response time of 2 seconds have been overlayed. The sampling rate has been chosen because of the number of data points forming a peak is 20 or less here. That way it is possible to visualize single data points during leading and trailing edges of the peaks (e.g. fig. 2.6 and fig. 2.7, upper images).

For a closer examination of possibly existing deviations specific scatter plots have been created. Each acquired data point produced by the CDS of one run has been plotted against the signal difference between the CDS and *Data Collector*. In order to reduce the number of shown data points in the scatter plot the time range from 80 to 120 seconds of the original chromatogram has been selected. This time range includes the third peak only. The scatter plots have been produced for all following comparisons that way.

Figure 2.7: Overlaid plots of *OpenLab ChemStation*[®] and *Data Collector* for 1200 HPLC series system

The plots of the raw data acquired with a sampling rate of 20 Hz (above) seem to be identical. In the scatter plot (below) of the third peak in the time range from 80 to 120 seconds there are three visible sections with increasing ranges of difference: 1-10 mAU, 10-100 mAU and >100 mAU.

Data Collector versus OpenLab ChemStation®

The figures 2.6 and 2.7 show the simultaneous data acquisition between *Data Collector* and *OpenLab ChemStation*® from *Agilent Technologies*. Figure 2.6 represents the 1260 series and fig. 2.7 the 1200 series HPLC system.

For both HPLC systems and both data acquisition software packages no visible deviations between the raw data plots were detected (upper images). The data points in the leading and trailing edges of the peaks seem to be supposedly the same. The scatter plots are similar and show an unit phenomenon. The order of magnitude of the differences increases whenever order of magnitude of the original signal increases. In detail a original data points with a magnitude of order of 10^0 and 10^1 lead to differences with a magnitude of order of 10^{-7} and 10^{-6} , respectively. This relationship is independent from the sampling rate and signal filtration and valid for all acquired chromatograms.

Figure 2.8: Overlaid plots of *Chromeleon*® and *Data Collector* for 1260 HPLC series system
 The data points in the raw data plot (above) seem to alternate between *Chromeleon*® and *Data Collector*. The appearance of the scatter plot (below) of the third peak in the time range from 80 to 120 seconds shows an increasing difference up to a maximum and a decreasing afterwards until a difference close to zero. The maximal deflection is lower for a greater sampling rate and lower response time.

Data Collector versus Chromeleon®

The simultaneous data acquisition between *Data Collector* and *Chromeleon*® from *Thermo Fischer* resulted into fig. 2.8 for the 1260 series and fig. 2.9 for the 1200 series HPLC system.

In contrast to the plots of the *Data Collector* and *OpenLab ChemStation*® there are clearly identifiable deviations in area of the peaks here (figs. 2.8 and 2.9, upper images). For 1260 series it looks like an alternating data acquisition between *Chromeleon* and *Data Collector* whereby for the 1200 series the signal differences are lower. The scatter plots provide a completely other appearance as seen in figs. 2.6 and 2.7 (lower images). First of all the maximal difference values are much higher compared to *Data Collector* versus *OpenLab ChemStation*®. The second difference is the formation of the data points. An increasing signal value results into an increasing absolute difference up to a maximum. After that maximum, which seems to correspond to the inflection point of the peak, the difference decreases while the signal is still getting higher until the difference is close to zero at the point of peak apex. For the data points in the falling part of the peak the differences are getting negative forming a minimum at the peak inflection point.

Figure 2.9: Overlaid plots of *Chromeleon*® and *Data Collector* for 1200 HPLC series system
The appearance of the chromatogram and scatter plot correspond to fig. 2.8 but the maximal deviation between *Chromeleon* and *Data Collector* is lower for 2.5 Hz sampling rate.

Even if the absolute, maximal difference for the 1200 system is lower, in both systems a greater sampling rate and lower response time influence the formation by decreasing the maximal deflection (figs. 2.8 and 2.9, lower image).

Data Collector versus ADC 35900 Dual Channel Interface

When using *OpenLab ChemStation*® for the data acquisition the provided data from the detector and from an external analog-to-digital converter (ADC) have been collected (fig. 2.2). The analog signal of the detector has been converted by the *ADC 35900 Dual Channel Interface* manufactured by *Agilent Technologies*. The figure 2.10 presents the overlaying raw data and scatter plots for the 1260 series HPLC system. In the overlaying chromatogram (upper image) a shift on the time axis can be seen very well. Therefore there are relative large signal differences between the ADC and the *Data Collector* at the same time in the scatter plot (fig. 2.10, lower image) and the differences at the peak apices are not close to zero. Using a higher sampling rate leads to a lower deviation.

Figure 2.10: Overlaid plots of analog-digital-converter and "Data Collector" for 1260 HPLC series system

Between the signal of "Data Collector" and the ADC is visible delay. This leads to high difference values in the scatter plot of the third peak. Additionally the difference is not close to zero at the maximal signal value.

In order to exclude the influence of the time delay the ADC data points have been shifted

Figure 2.11: Scatter plot of analog-digital-converter and *Data Collector* for 1260 HPLC series system after time shift

The scatter plot of the third peak deviates significantly compared to the scatter plot in fig. 2.10. The time shift for the 5 Hz ADC signal is 1.6 seconds and for 20 Hz 0.6 seconds.

by -1.6 seconds for the 5 Hz and -0.6 seconds for the 20 Hz signal. The given time values were chosen by determining the time difference between the maximum signal of the third peak (apex) in both chromatograms (ADC and *Data Collector*). The resulting scatter plot after the time shift of the ADC data is shown in fig. 2.11. There is still a deviation between the peak apices but the differences in the leading and trailing edges of the peaks decreased significantly for both sampling rates 5 and 20 Hz.

2.5.3 Preclusion of Data Manipulation by Licop Library via Data Sniffer

The own written *Data Collector* integrated an official library called *Licop* [27]. This library is used for the hardware communication to the HPLC systems manufactured by *Agilent Technologies*. In order to preclude any data manipulation by this library a simple experiment was performed whereby the detector data have been acquired by the *Data Collector* as usual but additionally data packages from the Ethernet connection have been caught (fig. 2.3) using the data sniffer tool *WireShark*.

The results of this experiment are listed in table A.1 on page 85 in the appendix section. That table contrasts the data points *Data Collector* stored in a text file and the data packages that were caught from the Ethernet connection during the run. Based on a document describing the meaning of the bytes [36] in all caught packages it has been found out that in a specific data section every four bytes are forming one signal value (a 32 bit data type, hexadecimal presentation in table A.1). These values are converted to decimal format then. The converted values are given as "counts" and in order to get them as "mAU" (milli absorbance unit) a conversion factor of 2097.153 counts per mAU has to be used [36]. Applying this factor to all data values caught from the Ethernet connection it is clearly evident that all data points acquired and stored by the *Data Collector* are completely equal to the caught

Ethernet connection values converted in "mAU" (table A.1).

2.5.4 In-Depth Analysis of Data Acquisition by Chromeleon via Data Sniffer

Based on the knowledge from the data sniffer experiment verifying data acquisition of the *Licop* library (section 2.5.3) the configuration has been reused to bypass the storage procedure of the CDS *Chromeleon*[®]. The experiment conditions are similar to the setup described in section 2.3 and fig. 2.3 except the fact that the CDS *Chromeleon*[®] is acquiring the data now instead of the *Data Collector*. The data from the Ethernet connection were caught by the data sniffer tool *WireShark* again.

In order to avoid incorrect settings several options available by *Chromeleon*[®] have been tried and compared to the raw data provided by the device. That means additionally to the configured sampling rate and signal filtration *Chromeleon*[®] allows a setup of incoming data handling. Two parameters called **step** and **average** are available for this. The step parameter is used to define a dynamic sampling rate using an internal algorithm or a fixed sampling rate with a specific step width in seconds for equidistant data. On the other hand, the average parameter defines the behavior of reducing the number of provided data points if the set step width is greater than the sampling period of the device⁴. If averaging is disabled every n-th data point of the original raw data will be stored. If it is enabled several original data points will be bunched to a new single one by computing the average. The main point here is that both parameters are working completely independent to the configurations within the device the data came from.

The 1200 series HPLC system was used for the experiments and has been configured to sample the data of the simple single infection with 2.5 Hz and a signal filtration of 2 s response time or with 20 Hz and 0.5 s, respectively. The step parameter of the CDS *Chromeleon*[®] was either set to a fix step width of 0.4 or 0.05 s which corresponds to a sampling rate of 2.5 or 20 Hz. The dynamic option of the step parameter was not tested due to the fact that the HPLC detector provided equidistant raw data and the goal was to acquire and store that kind of data.

The average parameter was tested in both ways, disabled and enabled. The results for a disabled averaging and both tested sampling rates are visible in fig. 2.12 (a and b). These plots show the overlapped chromatograms of the *Chromeleon*[®] export and the caught data from the Ethernet connection (network). Both plots have an apparent time shift between the signals. The difference here is that in fig. 2.12 (a) the signal of *Chromeleon*[®] has a delay related to the caught data. In fig. 2.12 (b) it is the other way round. Additionally the data sniffer caught more data points than the CDS stored.

For further investigations the time shift has been eliminated by shifting the caught data

⁴In other words the user set a lower sampling rate for the step parameter in *Chromeleon*[®] than the configured sampling rate of the HPLC detector module

Figure 2.12: In-Depth analysis of data storage by Chromeleon® and disabled averaging

The data provided by 1200 series HPLC series system acquired by *Chromeleon*® and additionally caught from the Ethernet connection by *WireShark* using two different sampling rate and signal filtration configurations (a and b). The average parameter of *Chromeleon*® is disabled here. The apparent time shifts have been normalized (c and d). The still existing deviations are plotted in (e) for both sampling rates.

+0.4 s in fig. 2.12 (a) and -0.4 s in fig. 2.12 (b) to align the retention times. That shifting results to the chromatograms shown in fig. 2.12 (c) and (d). Afterwards a deviation plot has been formed plotting the time against the differences of the signal values (fig. 2.12 (e)). Obviously, there are still deviations after the time shift but their extrema do not depend on the set sampling rate. The deviation plot of 2.5 Hz sampling rate has a wider range due to a peak broadening when using a more intense signal filtration [20].

A repetition of the whole experiment using an enabled averaging parameter lead to the chromatograms shown in fig. A.2 on page 90. In general the behavior is the similar. But the absolute deviation values for the 2.5 and 20 Hz samples (fig. A.2, e) are nearly three times greater compared to the setup with a disabled average parameter. The influence on the incoming raw data seems to be strengthened by the enabling the average parameter.

2.6 Discussion

A double entry method for the simultaneously data acquisition has been established and the data integrity of commercial Chromatography Data Systems CDSs have been investigated. A self written tool named *Data Collector* has been developed and used for that purpose. Based on the condition that two instances can communicate with the given HPLC system it is possible to acquire data twice and compare them afterwards. This extends the usual behavior that only one controller has access to the instrument and acquires its data.

2.6.1 Sampling Rate and Signal Filtration

For the 1260 series HPLC system manufactured by *Agilent Technologies* the peaks in chromatograms acquired by the *Data Collector* have been evaluated and compared for two experiment setups (section 2.5.1). It has been found out that changes in the sampling rate are coupled with the response time parameter for the signal filtration. That means that the filtration algorithm is embedded into the mainboard firmware of the given system and is independent from the controlling CDS during the data acquisition process. Relating to different terms describing a chromatographic system like retention time, peak efficiency and signal-to-noise ratio it has been determined that a lower sampling rate and higher response time lead to peak broadening, worse peak efficiency and reduced peak heights. On the other hand the signal-to-noise ratio increases significantly. Just the influences on retention times of the peaks are marginal. These results and the fact that no peak tailing phenomenons occurred for a lower sampling rate and higher response time correspond to already done investigations using a similar HPLC system and a sample containing two components [14].

Comparing two sampling rates configured for the ADC *35900 Dual Channel Interface* manufactured by *Agilent Technologies* has shown that without a signal filter the 10-fold lower sampling rate decreases the peak efficiency maximal by 10%. A 10-fold reduced sampling rate with signal filter of the HPLC detector unit lead to a maximal decreased peak efficiency of nearly 60%. But it has to be considered that the initial sampling rates are not the same.

The signal-to-noise ratio has been increased significantly using a lower sampling rate for the external ADC. That implies that a data bunching has been applied on the signal by the converter as suggested by [16, page 48].

2.6.2 Integrity and Comparability of Acquired Data by CDS

The data integrity between acquired and stored data of two frequently used commercial Chromatography Data Systems CDSs have been checked. The data comparison between the *Data Collector* and the CDS *OpenLab ChemStation*[®] from *Agilent Technologies* has shown that the data are completely equal except rounding deviations due to differences in the data accuracy. All data points from the *Data Collector* have a double precision accuracy whereby

for the exported data from *OpenLab ChemStation*[®] the accuracy has a floating precision. The increasing deviations visible in figs. 2.6 and 2.7 can be explained with the order of magnitude of the current data point. If the order of magnitude increases by one the number decimal places decreases by one. In fig. 2.7 (below) the acquired data sampled with 20 Hz exceeds a deviation of 10^{-5} because the third peak reaches a signal value of 100 mAU here. The reduced peak height by sampling with 2.5 Hz leads to peak values below 100 mAU here.

In contrast to *OpenLab ChemStation*[®] the deviations between the *Data Collector* and the CDS *Chromeleon*[®] from *Thermo Fischer* are significantly. This is an unexpected fact due to the *Licop* library which is used by the self written *Data Collector* but also by both official CDSs. This aspect has been determined on two observations. On the one hand the library file exists in the installation location of both CDSs. Furthermore the known commands that have been sent to HPLC have been caught by the sniffer *WireShark*. Based on these information and the gained experience that the *Licop* library has no influence on the incoming data all manipulations on the raw data are completely done after acquiring and during unknown, internal storage processes of *Chromeleon*[®]. No matter which step or average parameter settings have been configured within *Chromeleon*[®] it was not possible to receive the equal raw data as caught by the Ethernet connection sniffer. That means there is an algorithm like an interpolation applied to the raw data that can not be deactivated. For large peaks the resulting deviations may not be significant but for small peaks they are. That way it is not possible to compare data between two CDSs such as *OpenLab ChemStation*[®] and *Chromeleon*[®] as expected.

The ADC has a large time delay relating to the *Data Collector* presumable due to the additional provision of analog data from the HPLC detector unit to the external ADC *35900 Dual Channel Interface* itself. The effect of the time shift is decreased by using a higher sampling rate because the distances between each time point are reduced. A normalization by a manual time shift has shown that the deviations can be reduced significantly here.

All in all there are no manipulations on raw data provided by the detector unit that can be assigned to the *Licop* library and the *OpenLab ChemStation*[®]. The complete flow path from acquisition over storing until the export the raw data is completely traceable here. When *Chromeleon*[®] is used it has to be considered that the CDS itself becomes a not negligible parameter when validating and verifying a new experiment design. Furthermore it is important to know that it is possible to enter a step value corresponding to a sampling rate that is higher than the provided one. So not sampled but estimated data seem to be stored as raw data then.

For all further experiments in this thesis it is preferable to use raw data acquired by the *Data Collector* because here all steps from acquisition to storage are understood now.

2.7 Summary

The initial questions on page 9 can be answered as follows.

Conclusions

1. In practice and regulated environment one controller has access to the instrument only and acquires its data. Lock functionalities avoid manipulations of experiments by a second controller.
2. Two CDS have been tested. The raw data acquired and stored by *OpenLab ChemStation*[®] is traceable and not manipulated. A comparison between *OpenLab ChemStation*[®] and *Chromeleon*[®] would fail due to unknown processings on the data by *Chromeleon*[®] during the storage procedure.
3. A double entry method has been implemented successfully by a tool named *Data Collector* requesting as little as possible permission of the instrument using the *Licop* library. That allowed to acquire data simultaneously even if the system is locked by an official CDS.
4. The results of Wahab et al. [14] on sampling rate and response time relating to their HPLC system manufactured by *Agilent Technologies* could be reproduced with an isocratic standard containing four components.

3 Multi-Objective Optimized Signal Smoothing

After an excursion to the data storage procedure of commercial Chromatography Data Systems CDSs dealing with the data integrity and presentation of the double entry method for automatically acquired data this chapter focuses now on data only that will be generated and provided by the analytical instrument. As mentioned in chapter 1 it is desired to give recommendations to laboratory assistants or method developers relating to the detector sampling rate and signal filtration. For that purpose the following questions came up.

Questions

1. What are the recommendations in detail a user needs in order to optimize its method development in the regulated environment?
2. What are the prerequisites to give any recommendations?
3. What are the variable parameters and which kind of criteria are suitable for optimization of a chromatogram?
4. Are there differences in investigating the active substances or impurities of the pharmaceutical chromatograms?

3.1 Concept of Recommendation

In practice any laboratory assistants have to work with established accessories when they perform an analysis of an active substance in a medicine. Based on the knowledge of previous publications dealing with the sampling rate and signal filtration of the detection unit it has been shown that recommendations for these parameters are necessary but missing in the pharmacopoeias as mentioned in chapter 1. The purpose now is to find a way to check out the possible settings of the sampling rate and signal filtration and their influences on the chromatogram. By an appropriate evaluation the best combination of these parameters could be found and represented to the user then.

An approach possibly solving that problem and which will be investigated in this chap-

ter is visualized in fig. 3.1. An authorized user controlling the High Performance Liquid Chromatography (HPLC) system prepares an experiment as defined in one of the pharmacopoeias or any other standard operation procedure. The HPLC detector will be configured that way using the maximal possible sampling rate and the minimal possible influence of a signal filtration. If an Analog-To-Digital Converter (ADC) is available a data acquisition by that module is preferred because it allows a larger sampling rate and the signal filtration can be disabled here completely. The resulting chromatogram then functions as input for a procedure that will reduce the sampling rate and perform signal filtration algorithms on the signal in a structured way. Each change on the signal will be evaluated and in the end the best possible combination of sampling rate and signal filtration parameters will be provided to the user for further experiments using the given system.

Figure 3.1: Recommendation of Sampling Rate and Signal Filtration

Conceptional design how to determine the best possible parameters for sampling rate and signal filtration to give them as recommendations for further experiments.

Such a procedure optimizing the chromatographic signal in an automatic and suitable way is desired but not available yet [20]. That is why a new software tool has been developed during this thesis. It combines four features that should allow to improve the experimental results during a method development or of an already developed method by optimizing smoothing and denoising of acquired raw data. The tool got the name Smooth Signal Maker, abbreviated to *SmooSi Maker*. The sections 3.2 to 3.6 and 3.8 explain how the following four desired features of the tool have been realized during the development of *SmooSi Maker*.

1. Reduce the sampling rate in a most realistic way.
2. Perform filter algorithms that are established in the commercial and usually used detector units.

3. Find the optimal combination of sampling rate and signal filtration parameters by one or multiple evaluation criteria.
4. Consider the field of interest when optimizing chromatograms of pharmaceutical substances containing impurities.

3.2 Functionality 1 - Reduction of the Sampling Rate

As mentioned in section 1.3 there are several ways to reduce the sample rate. The most common way to reduce the number of data points is done by data bunching because this seems to be the most realistic approach due to the integrator used in Delta-Sigma analog digital converters installed in most data acquisition systems [14]. That means reducing the sample rate from for 50 Hz to 5 Hz leads to an averaging of 10 consecutive data points to a new single one. That behavior has also been determined for the ADC *35900 Dual Channel Interface* manufactured by *Agilente Technologies* in chapter 2. The baseline noise has been reduced appreciable by the usage of a lower sampling rate only (fig. A.1 in appendix). That implies an averaging of the data points. The data bunching method but also extracting each n -th data point to reduce the sample rate as an alternative have been implemented into the *SmooSi Maker*.

3.3 Functionality 2 - Estimation of Hardware Based and Performing of Firmware Based Signal Filtration

Filtration and denoising a given signal has been investigated for many years and so a wide range of filtration principles are available [20, 37]. But for this thesis not all of them are relevant. If it is desired to give advises to an user acquiring data with one of the established detector units only their kinds of filters matter. That is why the implemented and installed filters of established HPLC systems have been researched [14, 38, 39] and summarized in table 3.1.

Identifying the used signal filter is not as easy as expected. The kind of filter is rarely defined exactly in the corresponding manual of the detector unit. For one major vendor, *Shimadzu*, it was not even possible to find a hint which of the known filters is used or how their specific one works.

3.3.1 Hardware Based Filter

Hardware based filters are embedded to the hardware system and filter the generated signal physically. One established filter here is the so called RC filter using capacitive and ohmic resistances in series as shown in fig. 3.2. The intensity of the filter will be parameterized

Table 3.1: Kind of filters related to the HPLC vendors

This table lists all filters that could be researched and assigned to the HPLC vendors. Additionally, the filters are classified as physical filters which are installed into the detector as hardware components or as digital filters which algorithms are embedded into the motherboard firmware.

HPLC Manufacturer	Signal Filter	Filter Type
<i>Agilent Technologies</i>	Moving Average with Gaussian Kernel [14]	Digital Filter
<i>Thermo Fisher (Dionex)</i>	Moving Average or Savitzky-Golay [38]	Digital Filter
<i>Hitachi</i>	RC-Filter [14]	Physical Filter (Low Pass)
<i>Waters</i>	Hamming or RC-Filter [14, 39]	Digital Filter
		Physical Filter (Low Pass)

by the time constant ($\tau = R \cdot C$). Even if this kind of filter is a physical one an estimation equation for the RC filter exists [40] and will be used to simulate the filtration procedure within the *SmooSi Maker* tool (eq. (3.1)).

$$s_{out}(t) = \frac{\Delta t \cdot s_{in}(t) + RC \cdot s_{out}(t - \Delta t)}{\Delta t + RC} \quad (3.1)$$

Figure 3.2: Low Pass RC Filter Circuit [41]

Δt is the step width between two data points, RC corresponds to the time constant τ , s_{in} is the raw signal and s_{out} the resulting, filtered signal at the time point t . For the initial time point t_0 the filtered signal is equal to the raw signal.

3.3.2 Firmware Based Window Filter

All other listed filters belong to the digital filters and are implemented into the mainboard firmware of the detector devices. Additionally all the following filters have one similarity. They are window filters. That means a window with a specific size of N data points will pass the raw data and compute a filtered data point based on a weighting function applied on the data points within the window. The Window Size (WS) is always odd and the main parameter of all window filters.

Moving Average Filter

The most simple window filter is the moving average. Here all data points in the window have the same weighting as shown in eq. (3.2). That way the new data point will be computed by the arithmetic mean [14].

$$f_i = \frac{1}{WS} \cdot \sum_{j=-WS/2}^{WS/2} s_{i+j} \quad (3.2)$$

The term f is the filtered signal at position i , s is the raw signal and WS is the window size. The new signal data point will be computed by the original raw signal data point and preceding and following ones. That kind of filter is called centered moving average in detail.

Hamming Filter

The Hamming filter weights the data points within the window by the Richard Hamming function [20, 42] which definition is shown in eq. (3.3).

$$w_n = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{WS-1}\right) \quad (3.3)$$

Where w_n is filter weight at the zero-based index n and WS is the window size. Based on that definition it is possible to compute the centered filtered signal at position i by weighting the original raw data points (eq. (3.4)).

$$f_i = \sum_{n=0}^{WS-1} w_n \cdot s_{i-(WS/2)+n} \quad (3.4)$$

The filtered signal f at the position i is computed by weighting the raw signals s within the window using the Hamming weight w_n and sum them up.

Savitzky-Golay Filter

One of the traditional filter algorithms is Savitzky-Golay. Its approach is based on a noise reduction with minimal loss of relevant information when adjacent data points sharing similar information are averaged in a suitable way[37]. Suitable means that the data points in the window are weighted by a polynomial regression and averaged then. That leads to one additional parameter beside the WS : grade of polynomial. The weighting of each data point in the window depending on the grade of polynomial and WS can be computed [43]. The resulting coefficients can be used to calculate the filtered signal as like as for the Hamming filter (eq. (3.4)) replacing the term w_n by the corresponding Savitzky-Golay coefficient. The direct computation of the coefficients is very time consuming for large WS and grade of polynomial. That is why it is preferable to use precalculated coefficients and store them in a coefficient table. The implemented coefficient tables of the *SmooSi Maker* are prepared for

a maximal WS of 29 and grade of polynomial of 15.

Moving Average Filter with Gaussian Kernel

The last established filter is the centered moving average with Gaussian kernel. Corresponding to the name all data points within the filter window are weighted by the Gaussian density function of the standard normal distribution [14, 20]. The averaging of raw data here is shown in eq. (3.5).

$$f_i = \sum_{j=-WS/2}^{WS/2} s_{i+j} \cdot e^{-\pi \cdot \left(\frac{j}{SD}\right)^2} \quad (3.5)$$

The filtered signal f at the position i depends on the raw signal s , the WS and the standard deviation SD . That means beside the WS this filter has the standard deviation as an additional parameter which has to be considered.

3.4 Functionality 3 - Multi Objective Optimization

After implementing the reduction of sampling rate (section 3.2) and the performance of signal filtration algorithms (section 3.3) into the *SmooSi Maker* tool it is possible to simulate all the influences on the raw chromatograms by these detector unit parameters now. The follow-up question is how to evaluate these changes on the raw signal. How to perceive if the chromatogram is getting "better"?

Some recommendations were given by previous investigations on optimization of the filtration strength. One example is the estimation of optimal filter based on the largest negative chromatogram in the second derivative [20]. This iterative approach maximizes the absolute minimum of a peak in the second derivative which corresponds to the peak apex. An other approach is based on the Durbon Watson parameter [44]. This parameter is often used for regression data in order to determine whether correlations exist in the residuals. Applied on the chromatograms the differences between raw and filtered signal will be compared by that parameter and the convergence to a specific value is striven. Furthermore, a really similar approach to the Durbon Watson parameter is the smoothing index defined by Kumar [37]. Here a relative small value implies an intensive denoising and smooth profile of the signal.

All of these approaches concentrate on one evaluation value. Additionally, they mainly focus on the denoising of the signal only. When a chromatogram has been acquired in the regulated environment like the pharmacy all evaluating values are defined in the pharmacopoeias [3–5] (see fig. 1.1). During the method development the resolution is the most important variable that has to be optimized until all containing peaks are separated in a

suitable way. Based on the influences of the sampling rate and the signal filtration processes on the chromatograms which have been mentioned [14, 15, 20] four terms have been extracted that will be investigated: The peak width at half height ($W_{0.5}$), peak symmetry factor (A_s), signal-to-noise ratio (S/N) and peak resolution (R_s).

At this point there are multiple objectives now that have to be optimized. Therefore an optimization method dealing with multiple terms is required. Guantara reviewed two efficient and simple approaches that can be used for that purpose: the Pareto method and the scalarization method [45]. For this thesis the Pareto method has been chosen because for the scalarization method a suitable weighting of the objectives has to be found first and is not available yet. The Pareto method requires the information about minimizing or maximizing the individual objective only in order to find the so called non-dominated solutions (see section 3.4.1).

3.4.1 Pareto Optimization

The Pareto optimization belongs to the multi objective optimization (MOO) methods and keeps the objectives separated during the optimization process [45]. The only condition is to define if an objective has to be maximized or minimized. Then for each solution the objectives will be determined and related to a solution vector. The aim is to find the solution vectors that are non-dominated. Between two solutions s_1 and s_2 , s_1 dominates s_2 if

1. None of the objectives of solution s_1 is worse compared to the objectives of solution s_2 .
2. At least one objective of solution s_1 is better compared to solution s_2 .

By an iterative way comparing a new solution to all currently non-dominated solutions it is possible to form a so called Pareto Optimal Front (POF). A Pareto optimization for two objectives is illustrated in fig. 3.3. The blue solutions are dominated by at least one other solution whereby the green ones are non-dominated. The non-dominated solutions form a POF visualized as the green line. The utopia point corresponds to the ideal solution. The position of that point depends here on the definition that the peak width at half height and signal-to-noise ratio should be minimized and maximized, respectively.

Euclidean Distance

The Pareto Optimal Front (POF) usually consists of multiple solutions that can be considered as possible optima. Additionally it is possible to find the most suitable solution by calculating the Euclidean distance between the solutions of the POF and the utopia point [45]. Within the *SmoosSi Maker* tool the normalized Euclidean distance [46] has been implemented in order to select the optimal solution from the POF. The applied calculation to do so is summarized in eq. (3.6).

Figure 3.3: Multi objective optimization forming a Pareto Optimal Front (POF)

Several solutions with different peak widths at the half-height and signal-to-noise ratios. Four of the solutions are non-dominated (green dots) and form a POF (green line). Computing the minimal Euclidean distance between the POF solutions and the utopia point is one possible approach to select the best solution.

$$ED = \sqrt{\sum_{object} \left(\frac{v_{object} - v_{object}^*}{v_{norm}} \right)^2} \quad (3.6)$$

Where *object* is one of the criterion to optimize such as the peak width at half height, v is the object value of the current solution, v^* corresponds to the utopia point value of the object and v_{norm} is a normalization value that depends on the global minimal and maximal values of the current object in all determined solutions which belong to the POF.

3.4.2 Parameter Range of Interest

In section 3.3 the signal filtration algorithms and their parameters have been introduced. Each sampling rate, signal filter and the corresponding parameter such as the window size generates an other solution for the Pareto optimization. In order to check all suitable solutions the range of parameters has to be predefined because otherwise there will be infinite solutions even if they are nonsense.

Sampling Rate

For sampling rate it has been decided to reduce it starting from the raw data one. The reduction steps are oriented on the available sampling rates of the established detector units: 160, 100, 80, 40, 20, 10, 5, 2.5, 1.25 Hz. If the original raw has a sampling rate of for example 50 Hz, it will be reduced to 40 Hz first and continues from there.

Time Constant

The time constant parameters the physical RC filter. A larger time constant increases the strength of the filtration. So a value of 0 ms would lead to no filtration. It has been decided to start with a time constant that corresponds to the current sampling period ($1 / \text{sampling rate}$) and increasing it step wise even by this period. That means for a current sampling rate of 20 Hz with a corresponding sampling period of 50 ms the first tested time constant is 50 ms, too. Afterwards 100, 150, 200, etc. ms will be tested. The time constant will be increased up to a maximal value of $1/2 \cdot \text{peak width at the half height } (W_{0.5})$ because it has been determined that when the RC filter will be performed on an initial Gaussian peak its width exceeds an increasing of 25% then [14]. That is the upper border for the *SmooSiMaker* tool.

Window Size (WS)

For all firmware based filters the window size is the major or single parameter. The window contains at least 3 data points and the number is of points is odd. So the window size starts from 3 and will be increased by 2 for each performance. The maximal WS will be calculated by eq. (3.7).

$$WS_{max} = \frac{1}{2} \cdot \text{sampling rate} \cdot W_{0.5} \quad (3.7)$$

That calculation guarantees that no more than the half of data points forming a peak will be used for the filtration step. One exception exists. For the Savitzky-Golay filter the maximal WS is limited to 29 due to the prepared coefficient lists.

Grade of Polynomial

The Savitzky-Golay filter depends on the WS and the grade of polynomial. The grade of polynomial starts with 1 and increases by 1 until half of the current WS in order to avoid overfitting. Within the *SmooSiMaker* tool the coefficient tables have been prepared until a WS of 29 and a grad of polynomial of 15 due to the disproportionate runtime.

Standard Deviation

If a moving average filter with an Gaussian kernel is used standard deviation becomes a second parameter beside the WS. For optimization the standard deviation of the Gaussian starts from 0.1 and increases by 0.1 as well as done by Lytle et al. [20] up to half of WS because a higher standard deviation leads to weighting that is very close to an equal distribution.

Filter Repetition

Additional to all window filter parameters it is simply possible to repeat the filtration of the signal. For a RC filter this behavior is not realistic because a time constant is set and a repetition would increase the configured time constant. For the window filters each repetition increases the so called response time. That relation will be handled in section 3.6. Within *SmooSi Maker* it has been implemented to perform a filter repetition until the response time converges or exceeds a value equal to five times current sampling period. That maximal limit bases on the relation of sampling rate and response time in the detector unit manuals of *Agilent Technologies* [47, page 88].

Number of Peaks

The number of peaks do not belong to the variable parameters which will be tested during the optimization. It is a further hidden optimization criterion. Its aim is to avoid an optimization on the expense of detecting less peaks. So *SmooSi Maker* always try to maximize the number of peaks during the process in order to keep all of them.

3.5 Functionality 4 - Active Substance Analysis versus Purification Analysis

The last functionality of the *SmooSi Maker* is to distinguish between active substance and purification analyzes. Whereby concentrations of large peaks in the chromatogram are the main targets during active substance analyzes the impurities are of interest in purification analyzes as mentioned in section 1.1.2. Therefore three different types of peak considerations have been implemented.

1. Evaluate all identifiable peaks of the chromatogram to form a solution during the optimization procedure.
2. Evaluate the x largest peaks of the chromatogram only to form a solution during the optimization procedure. The peak heights are used to define the largest ones and x is given by the user.

3. Evaluate all identifiable peaks of the chromatogram which have a peak height lower equal to $p\%$ of the largest peak height to form a solution during the optimization procedure. The percentage p is given by the user.

3.6 Recommended Parameters - How to realize them?

Imagine an optimal solution using a specific sampling rate and signal filtration parameters have been found. The last step is then to overhand these recommendations to the user as shown in fig. 3.1. The problem now is that signal filtration parameters such as Window Size (WS), grade of polynomial and standard deviation are not available to set them for the real detector unit. Only two signal filtration parameters are used here: time constant and response time. The time constant is used for the RC filter and can be determined by the optimization procedure directly. But for detectors using the response time the definition of the American Society for Testing and Materials (ASTM) has been used [48]: The response time is defined as the time required for the signal to rise from 10% to 90% in response to an upward step function (fig. 3.4).

Figure 3.4: Response Time Determination

The upward step function visualizes the relation between signal raise and response time. Performing a signal filtration on the function increases the response time. A repetition of the filtration increases it once more.

That upward step function can be used now to convert the signal filtration parameter to response time. The filter will be performed on the step function and by a following simple linear interpolation between data points below and above 10% and 90% signal the time values of these signals will be determined. The difference between these time values leads to the response time then. If a time constant parameter is required even if the signal filtration parameter is not based on the RC filter the relation $response\ time = 2.2 \cdot time\ constant$ can be used[14].

3.7 Simulation of Chromatograms

Beside the acquisition of real chromatograms some investigations on the *SmooSi Maker* (section 3.8) and the *Persistence Filter* (chapter 4) will be done on simulated chromatograms. In order to make them reproducible the fundamental part of the HPLC simulator published by Boswell et al. [49] has been extracted. The final chromatogram construction equation is:

$$C_i(t) = \frac{W_i}{2\sqrt{\pi}\sigma_{t,i}F} \cdot e^{-\left(\frac{t-t_{R,i}}{2\sigma_{t,i}}\right)^2} \quad (3.8)$$

All the other calculations influencing the terms of eq. (3.8) are ignored because it is not necessary to simulate the whole HPLC experiment. The equation represents a function of the molarity C of compound i depending on the time t and is obviously a modified probability density function of the normal distribution. Beside the constants 2 and π there are the editable eluent flow F and the compound i specific number of moles W , peak width σ_t and retention time t_R . For any simulated chromatogram used in this thesis the parameters are listed in the appendix.

When n compounds are defined the total chromatogram $C_T(t)$ is calculated as the superposition of each compound then.

$$C_T(t) = \sum_{i=1}^n C_i(t) \quad (3.9)$$

For a more realistic chromatogram a baseline noise is missing. Boswell et al. add a random Gaussian noise to each data point for that purpose regulated by a standard deviation [49]. A comparison of that approach with a realistic chromatogram resulted in the conclusion that the deflections of the noise are sometimes too large because the noise calculation does not include previous, overlapping noise value. For that reason it has been decided to weight the noise calculation with the previous value. Have a look to eq. (3.10).

$$\text{Noise value} = N(t) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{px^2}{2\sigma^2}} & \text{if } t = 0; \\ \frac{1}{4} \left(3N(t - \Delta t) + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{px^2}{2\sigma^2}} \right) & \text{otherwise;} \end{cases} \quad (3.10)$$

with

$$\{px \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

The noise standard deviation of any used simulated chromatogram is also noticed in the appendix.

3.8 SmooSi Maker - Smooth Signal Maker

All aspects of the previous sections have been collected and implemented into a self written tool named *SmooSi Maker*. It is written in C# and requires at least the *Microsoft .NET 4.5* framework installed on Windows operation system. A release of the software with revision 0.9 is available on the enclosed CD of the physical version of this thesis.

Starting the release opens a screen as visible in fig. 3.5. The original raw data has to be converted into the comma separated value (*.csv) format and can be imported that way. Afterwards the user can enter some information about the raw data, define the time range for noise determination and may start the smoothing procedure after configuring the four functionalities of the *SmooSi Maker*: Sampling Reduction, Signal Filtration, Signal Optimization and Kind of Analysis.

Figure 3.5: *SmooSi Maker* Application Screen

The user can import raw data, configure and perform the signal smoothing and optimization and storing the result as an *SmooSi Maker* project.

3.8.1 Optimization of Chromatograms via SmooSi Maker

The application of *SmooSi Maker* has been tested on a simulated chromatogram containing a separation of three active substances and 18 impurities. The chromatogram is based on the publication of Vojta et al. presenting a new method to separate and determine impurities in Paracetamol, Codeine and Pitophenone [50]. The reconstruction of that chromatogram has been simulated using the chromatogram construction eq. (3.8) and the parameters of listed in table A.2 in the appendix.

The generated chromatogram has a sampling rate of 40 Hz. After importing the raw

data into the *SmooSi Maker* two smoothing procedures have been performed. Both of them reduces the sampling rate via data bunching, checks all available signal filtration algorithms and optimizes for the peaks width at half height, symmetry, signal-to-noise ratio and resolution. The only difference is that the first procedure optimized relating to the three active substances excluding the impurities and the second on was doing it vice versa.

3.8.2 Suitability of the Response Time Parameter

Even by the HPLC detector manuals and the definition of response time given by American Society for Testing and Materials (ASTM) [48] it is unknown how the response time parameter influences the parameters in the signal filtration algorithms. Therefore an experiment has been performed to compare the estimated filtration results of *SmooSi Maker* and the results of the 1200 series HPLC system after setting a reduced sampling rate and corresponding signal filtration response time.

For that reason a chromatogram separating Prednison, an anti-inflammatory and immunosuppressant listed in the European pharmacopoeia as monograph 01/2008:0354 [51, page 2791], and the usual used preservative Methyl paraben has been acquired. In order to simulate an experiment similar to a listed one in a pharmacopoeia the reference solution and two eluents defined in the monograph have been produced. Eluent 1 consists of 10 vol.% Acetonitrile, 20 vol.% Methanol and 70 vol.% HPLC-grade water. Eluent 2 is 100 vol.% Acetonitrile. For the reference solution a tablet containing 10 mg Prednison was solved in 100 ml Methanol and filtered afterwards to get a 100 µg/ml solution.

For the preservative Methyl paraben a 1 mg/ml solution has been produced by solving 10 mg Methyl paraben in 10 ml Methanol. The final test solution has been produced by mixing 2.5 ml Prednison solution and 0.25 ml Methyl paraben solution and fill the result up to 5 ml with eluent 1. That lead to a test solution containing 50 µg/ml Prednison and Methyl paraben in each case.

A HC-C18 Plus column supplied by *Agilent Technologies* was installed as stationary phase according to a reverse phase chromatography configuration. The column had a length of 250 mm, a diameter of 4.6 mm and a pore size of 5 µm. The chromatographic conditions are based on the configurations of the monograph of Prednison [51, page 2791] listed in table 3.2. Only the gradient setup has been adapted in order to reduce the run time and get the Prednison peak closer to the Methyl paraben one. All required instruments and materials listed in this section were provided by the company *AnaTox GmbH & Co. KG*. The injections have been repeated six times.

Table 3.2: Experimental conditions for separation of Prednison and Methyl paraben

The chromatographic conditions for the single run using the 1200 series HPLC system and a 50 µg/ml test solution of Prednison and Methyl paraben for the injection.

Parameter	1200 series HPLC system
Run Time [min]	12
Flow [ml/min]	1.5
Eluent Gradient [min : %B]	0.0 : 0 5.5 : 0 10.0 : 60 10.1 : 0
Injection Volume [µl]	20
Sampler Draw Speed [µl/min]	50
Sampler Eject Speed [µl/min]	200
Column Temperature [°C]	45 ± 0.5
Detector Signal [nm]	254
Detector Signal Bandwidth [nm]	4
Detector Reference [nm]	500
Detector Reference Bandwidth [nm]	100
Offset [%]	5
Sampling Rate [Hz]	20
Response Time [s]	0.125

3.9 Results

The *Smooth Signal Maker (SmooSi Maker)* has been implemented as a tool that could determine suitable sampling rates and signal filtration parameters of HPLC detector units in order to improve the detection and processing of peaks in chromatograms. Four functional aspects have been considered and implemented as mentioned in the sections 3.2 to 3.5.

3.9.1 Optimization of Signal Smoothing

The performance of the *SmooSi Maker* has been tested on a simulated chromatogram containing 21 peaks. Previous tests of the *SmooSi Maker* on three different chromatograms mainly focused on the Pareto optimization have been published in [52]. It has been shown that the Pareto optimization approach worked well and determining a suitable solution by the euclidean distance is possible. Since then the functionality of selecting the kind of analysis has been added to the *SmooSi Maker* and that is why the chromatogram shown in fig. 3.6 has been investigated twice.

Optimization Focused on Active Substances

Figure 3.6 contains three large peaks corresponding to Paracetamol, Codeine and Pitophe-none. These peaks have been investigated first by selecting the active substance analysis

Figure 3.6: Simulated Separation of Paracetamol, Codeine, Pitophenone and Impurities

With increasing retention time the three large peaks of Paracetamol, Codeine and Pitophenone are visible. Additionally there are 18 peaks which belong to impurities. The chromatogram has been simulated but is based on the publication of Vojta et al. [50]

within the *SmoosSi Maker* tool and setting the number of relevant peaks to 3. That way the software tool considered the three largest peaks only when evaluating the chromatogram. The results for the raw data and each performed signal filter is summarized in table 3.3. The solution with the bold font using the Hamming filter is the optimal solution overall after checking the Pareto Optimal Front (POF) of each filter.

It is apparent that for all performances of the filters and sampling rate reductions the symmetry and the signal-to-noise ratio have been improved compared to the raw data. On the other hand the peak width and resolution got worse. The Hamming filter has the least decrease of the peak width and resolution performance which seems to be the reason for being the optimal solution.

Optimization Focused on Impurities

The impurities of the chromatogram in fig. 3.6 have been considered in the second optimization procedure. Within *SmoosSi Maker* tool they have been defined by selecting purity analysis and setting the impurity height to lower than 50%. That means the software tool considered all peaks only that have a peak height lower than 50% of the largest one when evaluating the chromatogram. The results are listed in table 3.4. The solution with the bold font using a RC filter with alternate parameters here is the optimal solution overall after checking the Pareto Optimal Front (POF) of each filter.

As same as for the focus on the active substances the symmetry and the signal-to-noise

Table 3.3: Optimization related to active substances

The optimal parameters of all tested filters relating to the three active substances in the chromatogram of fig. 3.6 and the calculated terms determined by *SmoosSi Maker* are shown. Each term represents a sum of the considered peaks. The bold line is the overall optimum. Abbreviations: τ - Time Constant, WS - Window Size, GoP - Grade of Polynom, SD - Standard Deviation and Rep. - Repetition.

Parameter	$\sum$ Peak Width [ms]	$\sum 1\text{-Symmetry} $	$\sum$ S/N Ratio	$\sum$ Resolution
Raw Data 40 Hz	60814.677	2.036	158.126	61.809
No Filter 10 Hz	67365.752	0.039	191.62	55.793
RC Filter 40 Hz $\tau = 25$ ms	64615.879	0.024	179.424	58.082
Moving Average 20 Hz WS = 119	73028.039	0.546	834.733	48.199
Hamming 40 Hz WS = 3	61320.162	0.036	161.569	61.262
Savitzky-Golay 1.25 Hz WS = 9 GoP = 1	74536.076	0.682	998.41	46.508
Gaussian 20 Hz WS = 5 SD = 2.1 Rep. = 3	67372.31	0.225	194.744	55.832

ratio have been improved compared to the raw data whereby the peak width and resolution got worse. In contrast to the first experiment the overall, optimal solution is an additional one with alternate RC filter parameters.

3.9.2 Response Time Recommendation

After optimizing a given chromatogram by *SmoosSi Maker* it is desired to give an recommendation to the user for further experiments (see concept in fig. 3.1). Therefore an additional aspect to the smoothing procedure is to investigate the suitability of the conversion from sampling rate and signal filter parameter to response time which the can apply to his system. An chromatographic experiment where Prednison will be separated from the preservative Methyl paraben. The resulting chromatogram sampled with 20 Hz and 0.13 s response time is shown in fig. 3.7.

The experiment done with 20 Hz is used as raw data input for *SmoosSi Maker*. Then the smoothing procedure focussing on the two largest peaks in the chromatogram (via active substance analysis configuration) has been performed. The acquisition of a chromatogram and the following optimization by *SmoosSi Maker* has been repeated six times. Unfortunately one experiment failed as determined afterwards. That is why 5 experiments have been used

Table 3.4: Optimization related to impurities

The optimal parameters of all tested filters relating to the impurities in the chromatogram of fig. 3.6 and the calculated terms determined by *SmooSi Maker* are shown. Each term represents a sum of the considered peaks. The bold line is the overall optimum. Abbreviations: τ - Time Constant, WS - Window Size, GoP - Grade of Polynom and SD - Standard Deviation.

Parameter	$\sum$ Peak Width [ms]	$\sum 1\text{-Symmetry} $	$\sum$ S/N Ratio	$\sum$ Resolution
Raw Data 40 Hz	35758.353	3.641	122.646	1395.441
No Filter 20 Hz	59502.254	0.372	199.536	1373.661
RC Filter 40 Hz $\tau = 325$ ms	63336.615	1.406	226.014	1205.032
Moving Average 20 Hz WS = 3	61237.472	0.623	211.749	1362.172
Hamming 20 Hz WS = 5	61273.426	1.095	211.119	1361.274
Savitzky-Golay 20 Hz WS = 3 GoP = 1	61237.472	0.623	211.749	1362.172
Gaussian 20 Hz WS = 5 SD = 2.4	61327.056	1.131	211.49	1360.533
RC Filter 40 Hz $\tau = 225$ ms	62462.561	0.577	219.867	1265.954

for the evaluation. At the end *SmooSi Maker* gave the advice to set up 5 Hz or 10 Hz and corresponding filter parameter (table 3.5). For that reason the data acquisition has been repeated by setting up 5 and 10 Hz and 1 and 0.5 s response time, respectively, for the 1200 series HPLC system. The inspections of the chromatograms are summarized and listed in table 3.5.

The optimal solutions determined by *SmooSi Maker* are marked by a bold font in table 3.5. The corresponding evaluation of the acquired data from the HPLC system are listed one line above. Due to the deviation between the response time set up in the HPLC detector and the recommendations of *SmooSi Maker* additional intermediate results for 10 Hz have been extracted and added to the list. Here the response times are closer to the one of the detector. For 5 Hz none of the intermediate solutions came close enough to 1 s to investigate.

The deviation values in the table shows that at least one intermediate solution is much more closer to the evaluation values due to the closer response time.

Figure 3.7: Separation of Methyl Paraben and Prednison

The chromatogram has been generated by the 1200 series HPLC system and sampled by 20 Hz (0.125 s response time). Peak A belongs to Methyl paraben and peak B to Prednison.

3.10 Discussion

Following a concept show in fig. 3.1 to give recommendations relating to the sampling rate and signal filtration of HPLC detector units a new tool has been developed, the *SmoosSi Maker*. Several aspects of previous publications dealing with the topic of optimization and sampling rate/signal filtration have been considered and slipped into the development. The Pareto optimization has been implemented in order to evaluate the modified chromatogram from various perspectives.

Applied on a chromatogram containing 21 peaks, 3 large and 18 small ones, it has been shown that the implemented sampling reduction and signal filtration algorithms have influences on the raw data given as input. Not all criteria of the chromatograms like peak width or resolution have been improved by the optimization but this was not even expected. The definition of the euclidean distance allowed to find the solution where the criterion with the greatest possible improvement related to the others belongs to the optimum. Table 3.4 shows that the optimum within one filter (RC filter here) does not have to be the global optimum because the utopia point as well as the normalization value v^* in eq. (3.6) have a serious weighting here. The advantage of the given type of optimization is the low number of parameters to configure. If more control is desired weighting the criteria when calculating the euclidean distance could be appropriate but therefore much more experiences are required. Without any bias the *SmoosSi Maker* recommended an improved parameter setting as it was planned at the beginning of the development of that tool.

A not negligible disadvantage of *SmoosSi Maker* is that the smoothing process can take

Table 3.5: Application of recommendation in practice

The results for different calculated response times and the real chromatograms generated by the HPLC of 5 injections are summarized. The bold line is the optimum given by *SmoosSi Maker*. Abbreviations: τ - Time Constant, WS - Window Size, SD - Standard Deviation and Rep. - Repetition.

Parameter	Σ Peak Width [ms]	Devia- tion [%]	$\Sigma 1\text{-Sym-}$ metry	Devia- tion [%]	Σ S/N Ratio	Devia- tion [%]	Σ Reso- lution	Devia- tion [%]
HPLC Data 20 Hz, 0.13 s	9642.046 ± 53.506	/	0.148 ± 0.012	/	196.087 ± 8.359	/	42.542 ± 0.507	/
HPLC Data 10 Hz, 0.5 s	9694.569 ± 21.405	/	0.157 ± 0.022	/	232.319 ± 4.190	/	41.551 ± 0.409	/
Gaussian 10 Hz, 0.616 s WS = 9 SD = 4.3 Rep. = 2	9747.171 ± 52.909	0.54	0.256 ± 0.014	63.16	250.773 ± 11.374	7.94	42.063 ± 0.493	1.23
Gaussian 10 Hz, 0.528 s WS = 7 SD = 3.0 Rep. = 3	9722.720 ± 52.909	0.29	0.295 ± 0.014	87.76	240.339 ± 10.950	3.45	42.175 ± 0.493	1.50
Gaussian 10 Hz, 0.458 s WS = 9 SD = 4.5 Rep. = 1	9700.433 ± 53.013	0.06	0.135 ± 0.007	-14.09	230.447 ± 10.645	-0.81	42.276 ± 0.496	1.75
HPLC Data 5 Hz, 1 s	9843.816 ± 60.433	/	0.125 ± 0.046	/	330.282 ± 4.065	/	42.413 ± 0.382	/
Gaussian 5 Hz, 0.507 s WS = 3 SD = 1.4 Rep. = 3	9719.564 ± 55.138	-1.26	0.174 ± 0.016	38.76	238.502 ± 10.310	-27.79	42.184 ± 0.496	-0.54

a long time for the optimization when chromatograms with wide peaks and a huge number of data points are analyzed. That is because the range of investigated filter parameters depends on these aspects. The distinction between active substance and purity analyzes worked because the optimization results in table 3.3 and table 3.4 differ. As expected some of the solutions recommend sampling rates below 20 Hz for the active substances because there are enough data points forming the peak and the reduction of sampling rate do not lead to the loss of relevant information. But small peaks can disappear much earlier using a lower sampling rate. That is why no solution in table 3.4 fell below 20 Hz.

The response time as the desired result of the *SmoosSi Maker* for the user, who requires this parameter in order to apply it to his system, has been investigated and resulted to table 3.5. In summarize, when the optimized solutions recommends a response time that is not available for the given hardware the deviation between recommendation and results in practice increases. Therefore it makes sense to restrict the possible response times to user given ones as long as it is unknown how the manufacturer calculates the response time in detail.

3.11 Summary

The initial questions on page 33 can be answered as follows.

Conclusions

1. Relating to the pharmacopoeias the laboratory assistants get a list of all available parameters they have to configure except the sampling rate and signal filtration. Following the concept in fig. 3.1 the self written *SmooSi Maker* could help to find a good solution for these parameters. In other cases like the method development the concept helps to accelerate the optimization of the detector parameters.
2. Relating to the concept in fig. 3.1 one chromatogram with the maximal available sampling rate and a little as possible influence of the internal filter is suitable input for the *SmooSi Maker*.
3. The variable filter parameters are described and listed in section 3.4.2. The suitable optimization criteria depend on the kind of analysis. For impurity analyzes the signal-to-noise ratio is the main criterion whereby for analytical substances forming large peaks a good resolution is desired.
4. Mostly when optimizing related to impurities *SmooSi Maker* will recommend a high sampling rate in order to lose a peak. Additionally the peak terms like peak width at half height and signal-to-noise ratio have a stronger influence because these terms have low absolute values for small peaks.

4 Persistence Filter

The previous chapters have shown that the established signal filtration principles have a not negligible influence on the chromatogram and the containing peaks. For that reason new or other kind of filters have been reviewed like the wavelet analysis or chemometric approaches like principal component analysis (PCA) and multivariate curve resolution alternating least square (MCR-ALS) [37]. All of these techniques have one commonality, they try to denoise the signal and filter the significant information. The wavelet analysis is an improvement of the Short-time Fourier transform (STFM), requires various transformations and its parameters have to be optimized such as the parameters of the Savitzky-Golay algorithm mentioned in section 3.3.2 in order to get an adequate result. Both types of the chemometric approaches seem to perform a good denoise of the data but multiple acquisitions of chromatograms are necessary in order to determine the significant information and remove the irrelevant noise here [37].

The question now was if it is possible to remove the noise based on the given signal within one chromatogram and without any transformation to the frequency spectrum. Therefore a new kind of filter named *Persistence Filter* has been developed whereby the approach is an extension of the *Persistence 1D* algorithm published by Kozlov and Weinkauff [53]. This algorithm was designed and implemented for the extraction of local maxima and minima within an one dimensional function with one argument and one function value. The principle of the *Persistence Filter* will be explained in the next sections. In short the *Persistence Filter* is divided into two parts. First, the relevant information from the chromatogram will be extracted by locating all minima and maxima and filtering them by their persistence. Afterwards the resulting gaps in the filtered data will be reconstructed in a second step.

Questions

1. Is it possible to design a digital filter that focuses on denoising precisely?
2. Is it possible to improve the limit of detection and limit of quantification by such a filter denoising the signal?

4.1 Step 1: Filtration of not Relevant Data

The whole filtration procedure will be explained on a self-constructed example plot shown in fig. 4.1. The presented algorithm does not depend on the kind of data. That is why no axis legend is shown in the plot. In an usual chromatogram there would be a time signal relation. The significant information in the given plot are one small and one large peak. The large peak additionally contains two lower peaks, one in the leading and another in the trailing edge. Furthermore there is some baseline noise. The target of the *Persistence Filter* is now to extract all data points that represent the significant information of the plot whereby the noise is not included. All the other data points will be discarded. But how to define, determine and keep the relevant data?

Figure 4.1: Self-constructed example plot

The plot functions as an example the *Persistence Filter* will be presented on. Obviously information in the plot are a small and a large peak. The large peak additionally contains lower peaks at the leading and trailing edge. The baseline noise is target of the *Persistence Filter*.

The filtration bases on a combination of the *Persistence 1D* algorithm provided by Kozlov and Weinkauff [53], a previous inversion of the raw data and a definition of a threshold which will be used for filtering the resulting persistence values after the determination of local minimum-maximum pairs. The understanding of the *Persistence 1D* algorithm is important for comprehensibility of the whole *Persistence Filter*. For that reason the concept of *Persistence 1D* is summarized in fig. 4.2. Its main part of minima and maxima detection bases on the flooding watershed transformation introduced by Beucher in year 1979 [54]. The idea is to imagine the given diagram as a relief and to flood the complete relief beginning from the regional minima (fig. 4.2, 1, 2 and 4). When two flooded regions meet each other the resulting barrier is called watershed and represents a local maximum (fig. 4.2, 3 and

5). The *Persistence 1D* algorithm defines that a maximum-minimum-pair will be formed then between the maximum and minimum with the lower persistence. The persistence corresponds to the distance between the ordinate values of the minimum-maximum pair (fig. 4.2, 6). That way the global minimum will be left without a maximum paired with (fig. 4.2, red triangle).

Figure 4.2: Concept of the *Persistence 1D* algorithm based on the watershed transformation displayed on Caffeine spectrum

In the Caffeine spectrum from 190 to 300 nm wavelength there are two visible, local maxima and one minimum (excluding the minima at the boundaries). The principle of the watershed transform is shown from 1 to 6. The minima are flooded and when two regions meet each other a local maximum is formed (purple and orange triangles). That way minimum-maximum pairs are located and stored with their individual persistence ($\triangleq$ ordinate value distance). The global minimum displayed as red triangle has no paired maximum.

Within the *Persistence 1D* algorithm the flooded regions forming a water basin are called components. These components contain all data points left and right from the regional minimum until the corresponding maximum has been found. The idea now is to combine this behavior with a predefined threshold in order to filter components whose persistence value is below that threshold. All data points that belong to such a filtered component can be discarded then because they do not contain any relevant information. In the field of the chromatography we are mostly searching for positive peaks inside the chromatogram. For that reason it is more effective to invert the chromatogram before running the *Persistence 1D* algorithm. That way the resulting components will contain all data points of relevant peaks then. Figure 4.3 shows the inverted plot of fig. 4.1 and the red and green dots correspond

to the minimum-maximum pairs that will be found by the default watershed transformation with a threshold value of zero.

Figure 4.3: Inverted plot of fig. 4.1 and highlighted local maxima and minima

For the component determination the chromatogram will be inverted first. That way the positive peaks will form one corresponding component. For a better understanding of the components have a look to fig. 4.4. The red and green dots in the plot here mark all unfiltered minimum-maximum pairs (threshold=0).

If a threshold of five is used the number of minimum-maximum-pairs decreases to the significant ones as shown in fig. 4.4. All the data points that belong to the components of the significant pairs (red rectangle in fig. 4.4) and to the global minimum component will be kept. All other data points containing no relevant information are discarded. That means these data points will be completely removed from the set of data and not be replaced by a new value. The final result of the first *Persistence Filter* step will be a plot with small and large gaps between the kept data points shown as the red dots in fig. 4.5.

4.2 Step 2: Reconstruction of the Gaps using Laplace Interpolation

After the filtration of not relevant data points from the original raw data there are a low up to a huge number of gaps now. In order to perform further data processings like peak detection, peak integration and so on, which expect data with a fixed step width between the data points the gaps have to be filled. One simple but effective interpolation method is the Laplace interpolation used and adapted by Kozlov and Weinkauff [53] for reconstructing the signal after running the *Persistence 1D* algorithm. They solved the reconstruction problem using the Laplace operator and the quadratic programming method. In this thesis it has

Figure 4.4: Relevant maximum-minimum pairs and corresponding components

The number of minimum-maximum pairs are decreased compared to fig. 4.3 due to a threshold value of 5. The red rectangles visualize components of the minimum-maximum pairs. All data points that belong to a component of a not filtered pair will be stored.

been decided to use the iterative preconditioned biconjugate gradient method for that due to relative slow performance of the quadratic programming. For that purpose it is necessary to understand and transform the approach of the quadratic programming solver.

4.2.1 Quadratic Programming

Kozlov and Weinkauff implemented a solver for the reconstruction problem after extracting the local minima and maxima using the Laplace operator and the quadratic programming found in the source code of the *Persistence 1D* project [53]. In detail they use the *quadprog* functionality of MatLab [55]. This function performs a numerical optimization to compute the interpolation by the quadratic programming which is defined as follows [56]:

$$\min_x \frac{1}{2} x^T H x + f^T x \text{ such that } \begin{cases} A \cdot x & \leq b \\ A_{eq} \cdot x & = b_{eq} \\ lb \leq x & \leq ub \end{cases} \quad (4.1)$$

That optimization definition and the constraints use a lot of elements which will be described in the following.

x Vector containing all data points of the reconstructed signal by minimizing quadratic programming definition.

- H Quadratic objective term matrix of data points in the x vector. For laplacian reconstruction the discrete Laplace operator for one dimensional data is defined in that matrix (see eq. (4.2)). The discrete Laplace operator is an analog for the continuous Laplace operator in n -dimensional signals. The Laplace operator is also known as Laplacian and defined as sum of all non-mixed second partial derivatives of a differentiable function $\Delta = \sum_{k=1}^n \frac{\delta^2 f}{\delta x_k^2}$ [57].
- f Linear objective term vector of data points in the x vector. For laplacian reconstruction the weighted original data points are defined in that vector. See section 4.2.4 for more information. If no data weight is used all elements in that vector have the value zero.
- A Linear inequality constraints matrix. For laplacian reconstruction the elements will be defined such that way that either the constraint $x_i - x_{i+1} \leq 0$ (positive scope) or $x_{i+1} - x_i \leq 0$ (negative scope) are fulfilled. It depends on the data point x_i which could lie between a minimum and maximum and has a positive scope or vice versa. For that purpose all elements in the vector b have a value of zero.
- b Linear inequality constraints vector. See the description of matrix A . For laplacian reconstruction all elements in that vector have the value zero.
- A_{eq} Linear equality constraints matrix. For laplacian reconstruction all elements have value zero except data points that have not been discarded by the filter described in section 4.1. Their value is one.
- b_{eq} Linear equality constraints vector. For laplacian reconstruction it contains all original values of data points that have not been discarded by the filter described in section 4.1.
- lb Lower bounds vector defining the lowest value x_i may have. For laplacian reconstruction that constraint vector will not be used.
- ub Upper bounds vector defining the highest value x_i may have. For laplacian reconstruction that constraint vector will not be used.

The quadratic objective term matrix H containing the discrete Laplace operator kernel $[-1 \ 2 \ -1]$ has following form:

$$H = \text{Laplace Operator} = \Delta = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \quad (4.2)$$

Using all these definitions of the matrices and vectors and fill them with the corresponding information the Laplace interpolation can be applied. The resulting vector x contains all data points of the reconstructed data. The disadvantage of the quadratic programming is that it is relative slow and so the number of raw data is limited as mentioned in the source code of the *Persistence 1D* algorithm [53]. The sample data given by Kozlov and Weinkauff used to show the local minima and maxima location and following reconstruction contain a few data points only. But the developed *Persistence Filter* should be applied to chromatograms up to one hundred thousand data points. Therefore an other algorithm was needed to solve the Laplace operator in order to fill the gaps.

4.2.2 Iterative Preconditioned Biconjugate Gradient Method (PBCG)

The general definition of the Laplace interpolation is to find a function that satisfies the Laplace equation 4.3 in n dimensions for unknown data points (x_k is the argument of one dimension) and use the already measured data points as they are (see eq. (4.4)) [58].

$$0 = \Delta f(x_k) \quad (4.3)$$

That way for N data points there are N equations that have to be solved. But for an one dimensional function such as the chromatograms the equations themselves are simple and sparse. Therefore it has been decided to implement the iterative Conjugate Gradient Method (CG) [59] which can solve equation systems with the form $A \cdot x = b$. An efficient implementation for a variation of CG, the Preconditioned Biconjugate Gradient Method (PBCG), has been found in [60] and used for the reconstruction step of the *Persistence Filter*. The advantages of that variation are that it also deals with linear, but not necessarily positive-definite or symmetric, equations and it uses a preconditioner matrix in order to converge in fewer steps.

$$x_i = x_{(\text{measured})} \quad (4.4)$$

In order to transform Laplace transformation from quadratic programming to PBCG the quadratic objective term matrix H has to be converted to single equations. As mentioned before if a data point is known and not removed by the *Persistence Filter* the simple eq. (4.4) can be used. For each gap eq. (4.3) has to be computed. Three equations depending on the location of the gap can be extracted from the matrix H shown in eq. (4.2). The resulting Laplace interpolation equations are:

$$\begin{aligned}
0 &= x_i - x_{i+1} && \text{(left boundary)} \\
0 &= 2x_i - (x_{i+1} + x_{i-1}) && \text{(middle)} \\
0 &= x_i - x_{i-1} && \text{(right boundary)}
\end{aligned} \tag{4.5}$$

Based on these equations and the implemented PBCG the filtered raw data could be interpolated and lead to the reconstruct shown in fig. 4.5. Reconstructing the data by quadratic programming described in section 4.2.1 led to the same result.

Figure 4.5: Filtered raw data and laplacian reconstruction

The filtered raw data points using a threshold of five are presented as red dots. The gaps between the dots have been interpolated by the laplacian reconstruction (blue line).

4.2.3 Biharmonic and Triharmonic Reconstruction

Additional to the laplacian reconstruction Kozlov and Weinkauff presented adaptations to the Laplace operator in their *Persistence 1D* algorithm in order to reconstruct a C^1 - or C^2 -continuous smooth function. These terms describe the parametric continuity of the junctions between the original raw data and the reconstruct [61]. Both types are at least C^0 -continuous. That means the values of their junctions match. If their first derivatives also match at the junctions, they are called C^1 -continuous. A reconstruction of such a function can be reached by using the biharmonic operator instead of the laplacian one [62] shown in eq. (4.6).

$$H = \text{Biharmonic Operator} = \Delta^2 \tag{4.6}$$

Extracting the equations from the resulting quadratic objective term matrix H leads to the following five equations which can be used for the PBCG.

$$\begin{aligned}
0 &= 2x_i - 3x_{i+1} + x_{i+2} && \text{(left boundary)} \\
0 &= 6x_i - 3x_{i-1} - 4x_{i+1} + 2x_{i+2} && \text{(next to left boundary)} \\
0 &= 6x_i + x_{i-2} - 4(x_{i-1} + x_{i+1}) + x_{i+2} && \text{(middle)} \\
0 &= 6x_i - 3x_{i+1} - 4x_{i-1} + 2x_{i-2} && \text{(next to right boundary)} \\
0 &= 2x_i - 3x_{i-1} + x_{i-2} && \text{(right boundary)}
\end{aligned} \tag{4.7}$$

If it is desired that the junctions between the raw data and the reconstruct, their first and their second derivatives should be C^2 -continuous the triharmonic operator will be used [62].

$$H = \text{Triharmonic Operator} = \Delta^3 \tag{4.8}$$

Here seven possible locations are possible for the unknown data points.

$$\begin{aligned}
0 &= 5x_i - 9x_{i+1} + 5x_{i+2} - x_{i+3} && \text{(left boundary)} \\
0 &= 19x_i - 9x_{i-1} - 15x_{i+1} + 6x_{i+2} - x_{i+3} && \text{(next to left boundary)} \\
0 &= 20x_i + 5x_{i-2} - 15(x_{i-1} + x_{i+1}) + 6x_{i+2} - x_{i+3} && \text{(two apart from left boundary)} \\
0 &= 20x_i - x_{i-3} + 6x_{i-2} - 15(x_{i-1} + x_{i+1}) + 6x_{i+2} - x_{i+3} && \text{(middle)} \\
0 &= 20x_i + 5x_{i+2} - 15(x_{i+1} + x_{i-1}) + 6x_{i-2} - x_{i-3} && \text{(two apart from right boundary)} \\
0 &= 19x_i - 9x_{i+1} - 15x_{i-1} + 6x_{i-2} - x_{i-3} && \text{(next to right boundary)} \\
0 &= 5x_i - 9x_{i-1} + 5x_{i-2} - x_{i-3} && \text{(right boundary)}
\end{aligned} \tag{4.9}$$

All these cases for the reconstruction type (laplacian, biharmonic and triharmonic) and the corresponding equations have been implemented and applied to the result of the *Persistence Filter*. A comparison of the types is visualized in fig. 4.6. All types of reconstruction contain the significant data filtered by the *Persistence Filter*. That way there is no visible noise in the reconstructs compared to the raw data. But for the biharmonic and triharmonic reconstructs there are additional local extrema that have not been there before. One example is the maximum near the value 100 of the abscissa in the triharmonic plot. These possible unwanted additional extrema have been already mentioned by Kozlov et al. [63]. Here the quadratic programming method has an advantage against the PBCG. The Linear inequality constraints matrix A and vector b avoid additional extrema due to the monotonicity that can be defined for the unknown data points. These constraints are not available for PBCG

but is possible to include the original raw data in a weighted way to address this issue. This aspect is handled in section 4.2.4.

Figure 4.6: Comparison laplacian, biharmonic, triharmonic reconstruction

The reconstruction of the filtered raw data using a the biharmonic or triharmonic operator lead to a different reconstruct than the laplacian one. All three types contain the significant information of the raw data plot but there are additional extrema in the biharmonic and triharmonic plots that did not exist in the raw data.

4.2.4 Weighted Inclusion of Original Data into the Reconstruction

One further option that Kozlov and Weinkauff integrated into their reconstruction following the *Persistence 1D* algorithm is the ability to include the original data in a weighted way [53]. Therefore they defined a new parameter called Data Weight (DW) and used this parameter to edit two elements of the quadratic programming optimization (eq. (4.1)): The quadratic objective term matrix H and the linear objective term vector f .

If no data weight is used every element in the vector f has the value zero. Otherwise if a Data Weight is set between $0 < DW \leq 1$ the weighted linear objective term vector f_{DW} is calculated as following using the original raw data vector y .

$$f_{DW} = -2 \cdot DW \cdot (y_1, y_2, \dots, y_n) \quad (4.10)$$

For the quadratic objective term matrix the data weight influences the diagonal by adding the double of the DW to it (eq. (4.11)). That way the influence of the data point x_i within the laplacian, biharmonic or triharmonic kernel will be increased compared to the surrounding ones.

$$H_{DW} = H + 2 \cdot \begin{pmatrix} DW & 0 & 0 & 0 & 0 & 0 \\ 0 & DW & 0 & 0 & 0 & 0 \\ 0 & 0 & DW & 0 & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & DW & 0 \\ 0 & 0 & 0 & 0 & 0 & DW \end{pmatrix} \quad (4.11)$$

Equation (4.11) can be adapted to the Preconditioned Biconjugate Gradient Method (PBCG) approach of the reconstruction easily. For all equations in eqs. (4.5), (4.7) and (4.9) the factor of the data point x_i will be increased by $2 \cdot DW$. For example the equation for an unknown middle point in the laplacian reconstruction (see eq. (4.5)) is changed to the following then:

$$0 = (2 + 2 \cdot DW)x_i - (x_{i+1} + x_{i-1}) \quad (4.12)$$

After that the real raw data values are still missing. For that purpose the general Laplace equation (4.3) was combined with eq. (4.10) to form a changed condition:

$$-2 \cdot DW \cdot y_i = \Delta f(x_i) \quad (4.13)$$

Several checks on the interpolation using PBCG compared to the results of the quadratic programming have shown that they are most similar when the factor 2 is used instead of -2 forming eq. (4.14).

$$2 \cdot DW \cdot y_i = \Delta f(x_i) \quad (4.14)$$

The results of the resconstructions between quadratic programming and PBCG are really similar for low DW like 10^{-6} and 10^{-5} . Increasing DW lead to growing deviations but this is not a negative aspect as long as the reconstruct alignes more and more to the raw data when the DW raises. Figure 4.7 shows a similar plot to fig. 4.6 using a DW of 10^{-2} here. The main differences are the eliminated or at least reduced additional extrema produced by the biharmonic and triharmonic reconstruction types. Further investigations on the DW are handled in section 4.4.1.

Figure 4.7: Comparison laplacian, biharmonic, triharmonic reconstruction using a Data Weight of 0.01

Compared to fig. 4.6 the additional extrema of the biharmonic and triharmonic reconstructs have been reduced or eliminated.

4.3 Editable Parameters

In sections 4.1 and 4.2 a new kind of filter, the *Persistence Filter*, has been introduced. In sum the filter removes not relevant data and reconstructs the occurred gaps by Laplace interpolation using the Preconditioned Biconjugate Gradient Method (PBCG). Three parameters are available to edit the behaviour of the filter and which have to be investigated: **threshold (TH)**, **data weight (DW)** and **type of reconstruction** (laplacian, biharmonic, triharmonic).

The threshold defines which kind of signal changes are significant and which can be removed due to missing relevant information. The data weight aligns the reconstruct to the original raw data before the filtration. Kozlov and Weinkauff recommended a data weight of 10^{-6} [53]. For laplacian reconstructions the data weight will be investigated logarithmically from 1 to 10^{-6} . Additionally a data weight of zero (no influence of the raw data) will be checked. On the basis of fig. 4.7 the data weight will not be lower than 10^{-2} for the biharmonic and triharmonic reconstruction due to quadratic or cubic overfitting within larger gaps which could lead to additional extrema that are not available in the original raw data.

4.4 Results

The new filter algorithm named *Persistence Filter* was introduced. Its effects have been checked. Especially the influences of the parameters **Threshold** and **Data Weight (DW)** on the filtration and reconstruction processes are fields of interest. Furthermore, it will be checked if there are significant differences when applying the filter on large and small peaks. Therefore the described *Persistence Filter* algorithm has been implemented into the *SmoosMaker* (section 3.8) as a new selectable filter.

4.4.1 Influence of Data Weight (DW) Value

The first parameter of the *Persistence Filter* that has been investigated is the Data Weight (DW) in order to get a feeling about the influence of that variable. A simulated chromatogram using the constructor equation 3.8 and the parameters listed in table A.3 is shown in fig. 4.8. The left image shows that a data weight of 1 does not completely cover the original raw data (black) but the reconstruct follows all deflections (red). The threshold has been set to the height of the large peak here. That is why in the right image both small peaks (fused with and separated from the large one) have been completely removed from the chromatogram when a DW of 0 is used. Increasing the DW now logarithmically shows that the peaks but also the noise will be reconstructed. A DW of 0.01 and greater reconstructs the complete height of the two small peaks here.

Figure 4.8: Persistence filtration and reconstruction depending on DW

A simulated chromatogram containing a large and a small peak has been filtered by the *Persistence Filter* using a threshold equal to the height of the large peak. Without any DW the small peak will be completely removed from the chromatogram (red line on right image). Increasing the DW logarithmically reconstructs the small peak as well as the baseline noise. The red line on the left image follows all signal changes of the black raw data even when the amplitudes are not completely equal.

4.4.2 Application of Persistence Filter on Large Peaks

The *Persistence Filter* has been applied on the chromatogram of the isocratic test standard which contains four substances producing relative large peaks. The target is to check the calculable properties now that will be significantly influenced by the filter. All implemented types of reconstruction and DWs have been tested on the given chromatogram and the results are listed in table 4.1. The peak width at half height, the symmetry, signal-to-noise ratio of the raw chromatogram of fig. 4.9 are shown in the first line of the table and are used as reference to evaluate the changes after filtration. All the computed terms correspond to the sum of all four peaks. For all kind of reconstructions and DWs the terms have been calculated followed by determination of the percentage deviation relative to the raw chromatogram. In short the *Persistence Filter* has an significant influence to the signal-to-noise only because for all other evaluation terms the deviations are below 1%.

Figure 4.9: Isocratic Test Standard

The isocratic test standard contains the four nonpolar substances Dimethyl phthalate (A), Diethyl phthalate (B), Biphenyl (C) and o-Terphenyl (D) forming four large peaks. The chromatogram has been acquired with a sampling rate of 80 Hz.

Table 4.1: Persistence Filtration of Isocratic Test Standard

The threshold of the *Persistence Filter* is fixed to 0.0757 mAU and has been determined by the peak to peak noise of the noise in the time range above 200 seconds (not shown in fig. 4.9). The first column describes the used reconstruction type and the Data Weight (DW). All other columns contain the calculated chromatographic terms as sum of all four peaks. A percentage deviation relative to the raw chromatogram is also listed to each term.

Parameter	Peak Width [ms]	Deviation [%]	1-Symmetry	Deviation [%]	S/N Ratio	Deviation [%]	Resolution	Deviation [%]
Raw Data	13493.066	/	2.716	/	13658.328	/	36.044	/
<u>Laplacian</u>								
DW = 1	13494.292	0.009	2.716	0	14196.653	3.941	36.039	-0.014
DW = 10^{-1}	13494.296	0.009	2.716	0.000	15748.946	15.307	36.039	-0.014
DW = 10^{-2}	13494.314	0.009	2.716	0.000	18256.793	33.668	36.039	-0.014
DW = 10^{-3}	13494.316	0.009	2.716	0.000	20801.261	52.297	36.039	-0.014
DW = 10^{-4}	13505.077	0.089	2.732	0.589	29862.357	118.638	36.020	-0.067
DW = 10^{-5}	13505.077	0.089	2.732	0.589	31235.710	128.694	36.020	-0.067
DW = 10^{-6}	13505.077	0.089	2.732	0.589	31283.528	129.044	36.020	-0.067
DW = 0	13505.077	0.089	2.732	0.589	31288.839	129.082	36.020	-0.067
<u>Biharmonic</u>								
DW = 1	13494.300	0.009	2.716	0.000	14092.768	3.181	36.039	-0.014
DW = 10^{-1}	13494.299	0.009	2.716	0.000	14772.631	8.158	36.039	-0.014
DW = 10^{-2}	13494.310	0.009	2.716	0.000	15628.924	14.428	36.039	-0.014
<u>Triharmonic</u>								
DW = 1	13494.318	0.009	2.716	0.000	14100.001	3.234	36.039	-0.014
DW = 10^{-1}	13494.362	0.010	2.716	0.000	14417.483	5.558	36.039	-0.014
DW = 10^{-2}	13494.347	0.009	2.716	0.000	14789.932	8.285	36.039	-0.014

4.4.3 Threshold versus Data Weight when Applying Persistence Filter on Small Peaks

The influence of the *Persistence Filter* on large peaks is limited to the signal-to-noise ratio. But what about small peaks? An successful increasing of the signal-to-noise ratio would have a direct influence on the limit of detection here. If peaks near the baseline noise can be detected and integrated this would be a great improvement of the chromatogram. For that reason the influence of the filter on small peaks has been tested by editing the threshold and the DW parameters. As shown in fig. 4.8 a small peak could be reconstructed not only by a suitable threshold but also by a DW value greater equal 0.01.

Figure 4.10: Application of Persistence Filter on small peak

Two approaches dealing with small peaks are presented. On the left side the used threshold excludes the small peak which will be reconstructed by an applied DW only. On the right side an suitable threshold is used to keep the small peak even if no DW is used.

Figure 4.10 shows the four tested scenarios. The parameters of the simulated chromatogram are listed in table A.4. Firstly a threshold exceeding the small peak is used combining with a DW of 0 and 0.01. Then the threshold has been reduced to 0.006 in

order to include the small peak independent of the DW value. In all cases the type of reconstruction is the Laplacian one because here a DW of 0 is available.

Table 4.2: Filter influences on large peak

The peak detection and evaluation terms of the large peak after applying the *Persistence Filter*. Abbreviations: TH - Threshold, DW - Data Weight.

Chromatogram	Liftoff Time [s]	Retention Time [s]	Touchdown Time [s]	Peak Width [s]	Symmetry Factor	S/N Ratio
Raw Data w/o Noise	9.68	12.00	13.88	1.165	0.999	$1.54 \cdot 10^{46}$
Raw Data	10.22	12.00	13.67	1.163	0.999	1205.479
TH 1, DW=0.00	9.98	12.00	13.70	1.164	0.998	29091.686
TH 1, DW=0.01	9.98	12.00	13.70	1.164	0.998	1604.786
TH 0.006, DW=0	9.98	12.00	13.70	1.164	0.998	17962.563
TH 0.006, DW=0.01	9.98	12.00	13.70	1.164	0.998	1605.613

Related to the large peak in fig. 4.10 any affect on the chromatogram by threshold or DW lead to marginal deviations in peak detection and evaluation as shown in table 4.2. But for the small peak the usage of the suitable threshold of 0.006 is more accurate due to more similar values with the ideal raw data without noise. See table 4.3 for that.

Table 4.3: Filter influences on small peak

The peak detection and evaluation terms of the small peak after applying the *Persistence Filter*. Abbreviations: TH - Threshold, DW - Data Weight.

Chromatogram	Liftoff Time [s]	Retention Time [s]	Touchdown Time [s]	Peak Width [s]	Symmetry Factor	S/N Ratio
Raw Data w/o Noise	26.70	30.00	32.68	1.665	0.999	$5.35 \cdot 10^{43}$
Raw Data	29.39	30.01	31.15	0.848	1.435	2.758
TH 1, DW=0.01	28.61	29.99	31.58	2.174	1.079	4.978
TH 0.006, DW=0	27.23	30.01	31.58	1.287	0.832	75.132
TH 0.006, DW=0.01	28.63	30.01	30.94	1.044	0.832	5.018

4.5 Discussion

A new kind of filter has been developed and implemented. It is a combination of excluding data without any information and reconstruction of resulting gaps. Three parameters, the threshold, Data Weight (DW) and type of reconstruction influence the mechanism of that filter. The threshold allows to define which information in the given chromatogram is relevant and which is not. Increasing the DW include the original raw data to the final result of the *Persistence Filter*. A DW of 0.01 seems to be enough to reconstruct the heights of peaks predominantly. The type of reconstruction has only a little influence on the filtration due to the fact that for biharmonic and triharmonic reconstruction a DW below 0.01 is not suitable because of possible unwanted information that are not available in the original raw data. But for optimization procedures even the small differences between laplacian, biharmonic and triharmonic could lead to an optimum for a given chromatogram.

One of the major advantages of the Preconditioned Biconjugate Gradient Method (PBCG) against the quadratic programming approach is the fast performance when reconstructing the gaps which occur after the persistence filtration. That way it is possible to apply the Laplace reconstruction on chromatograms with a huge number of data points. The influences of the whole *Persistence Filter* algorithm have been checked for large peaks and it has been shown that changes in peak distortion (peak width and symmetry) and resolution are marginal. Their deviations related to the original data are below 1%. That means there is no significant improvement and deterioration on these terms possible when using the *Persistence Filter*. Only the signal-to-noise ratio is affected significantly. For the given chromatogram of the isocratic test sample an improvement up to 129% was possible. The signal-to-noise ratio and the related denoising is affected directly by the filter.

If the signal-to-noise ratio of large peaks have a large value even before the *Persistence Filter* has been applied there is not so much improvement on the chromatograms because the peaks are detectable in either case. But for small peaks the signal-to-noise ratio has a considerable influence on the limit of detection and limit of quantification. Both terms depend on the standard deviation of the baseline noise. Decreasing their values is desired because it allows to detect and quantify smaller peaks. Table 4.3 implied that using the *Persistence Filter* with a suitable threshold and a DW as low as possible seems to be the best way to reduce baseline noise around small peaks in order to increase signal-to-noise ratio of them significantly. That allows to detect and quantify these peaks more precisely.

4.6 Summary

The initial questions on page 54 can be answered as follows.

Conclusions

1. Applied on a chromatogram of an isocratic standard containing four large peaks it has been investigated that a new kind of filter, the presented *Persistence Filter*, affects only the signal-to-noise ratio of these peaks significantly. All other terms are influenced marginally.
2. The presentation of the relation between LOD and LOQ and the signal-to-noise ratio in section 1.3 makes it clear that low detection and quantification limits require signal-to-noise ratios consistent with the method precision and accuracy requirements. That means the only way to lower the LOD and LOQ is to reduce the noise, increase the signal or both [17]. The presented *Persistence Filter* deals with the first aspect: Reduce the noise. Shown on a peak very next to the baseline noise in fig. 4.10 it is possible to detect that peak and increase its signal-to-noise ratio significantly after applying the *Persistence Filter* with suitable threshold and data weight parameters. That allows to detect and quantify these peaks more precisely.

5 Conclusion and Outlook

The chapters and sections of the presented thesis led through the fields of data acquisition and processing from various perspectives. The path of the raw data, that have been generated by a detector unit of HPLC system and provided to a Chromatography Data System (CDS) which stores the incoming raw data to make it available for possibly desired export, has been examined.

It has been found out that not every available software dealing with chromatographic data behaves as expected. The commercial CDSs *OpenLab ChemStation*[®] and the internal used *Licop* library acquire provided raw data as they are and store them without any manipulation. The integrity of the raw data is fulfilled here because by the establishment of a double entry method using a new self-written tool named *Data Collector* a subsequent comparison of simultaneously acquired raw data and the CDS export was possible and confirmed the complete equality here. A second tested CDS called *Chromeleon*[®] exports manipulated data and even after several tests and configurations it was not possible to avoid these manipulations. As long as the deviations are not significant it may not matter but the clear definition of data integrity and the including traceability of the data are violated. That aspect has been considered for all other tests in this thesis by using the *Data Collector* for necessary acquisitions only.

In view of the field of pharmacy the pharmacopoeias have been examined and a deficit in their specification has been discovered. Although all configurable parameters of chromatographic tests which belong to the monographs within the pharmacopoeias are listed and well defined, two detector unit parameters are missing here: sampling rate and signal filtration. Their essential influences on chromatograms have been shown by several publications and a few reviews in this thesis (section 2.5.1). Dealing with these parameters was the main motivation of the thesis and the task was to develop a concept that may allow to determine suitable configurations of the parameters which can be recommended. A concept has been conceived and all required aspect have been implemented into one self-written smooth signal maker named as *SmooSi Maker*. Based on a very few number of input parameters *SmooSi Maker* optimizes a given chromatogram depending on user selections. That means there are different solutions when focusing on impurities instead of active substances forming larger peaks within a chromatogram. Furthermore the recommended parameters at the end of the smoothing procedure are converted to applicable ones like the response time as filtration parameter in most of the detector units of HPLCs. That allows to accelerate the method

development and harmonization and update procedures of the pharmacopoeias. Further developments on the *SmooSi Maker* are necessary in order to improve the runtime and including user weightings related to the optimization criteria.

During the development of *SmooSi Maker* a concept for a new kind of filter arose. It combines the complete removal of irrelevant information within a chromatogram and a following reconstruction of the resulted gaps. At the end the *Persistence Filter* has been developed which allows to increase the signal-to-noise ratio significantly whereby other terms of the peaks like width and symmetry remain. Furthermore the suitability to increase the precision of peak detection of small peak close to the baseline noise has been shown. This is a main step to improve the terms Limit of Detection (LOD) and Limit of Quantification (LOQ) which are related to signal-to-noise.

Publications

Conference Contributions

1. David T. Marehn, Detlef Wilhelm, Heike Pospisil, Roberto Pizzoferrato, "Reliability of data detection and processing approaches in chromatography", German Conference on Bioinformatics, Tuebingen, Germany, 2017.
2. David T. Marehn, Detlef Wilhelm, Heike Pospisil and Roberto Pizzoferrato, "Double entry method for the verification of data a chromatography data system receives," Sensors and Measuring Systems, 19th ITG/GMA-Symposium, Nuremberg, Germany, 2018, pp. 1-4.
3. David T. Marehn, Detlef Wilhelm, Heike Pospisil and Roberto Pizzoferrato, "Proving the Preclusion of Data Manipulation Using Parallel Data Acquisition in Chromatography", Materials Science Forum, Vol. 941, 2018, pp. 2390-2394.
4. David T. Marehn, Detlef Wilhelm, Heike Pospisil and Roberto Pizzoferrato, "Recommendations for optimal sampling and filter rates in liquid chromatography", Sensors and Measuring Systems, 20th ITG/GMA-Symposium, Nuremberg, Germany, 2019, pp. 1-6.

Journals

1. David T. Marehn, Detlef Wilhelm, Heike Pospisil and Roberto Pizzoferrato, "Double entry method for the verification of data a chromatography data system receives", J. Sens. Sens. Syst., 8, Extended Special Issue of the corresponding Conference Contribution, 2019, pp. 207-214, <https://doi.org/10.5194/jsss-8-207-2019>.

Bibliography

- [1] Michail Semonowitsch Tswett. “On a new category of adsorption phenomena and their application to biochemical analysis”. In: *Lecture at the March 8* (1903), pp. 20–39.
- [2] Chromatography Today. *5 Uses of Chromatography in Everyday Life Chromatography Today*. Journal. 2014-12. URL: <https://www.chromatographytoday.com/news/industrial-news/39/breaking-news/5-uses-of-chromatography-in-everyday-life/32639> (visited on 2019-05-12).
- [3] The United States Pharmacopeial Convention. “<621> Chromatography”. In: *The United States Pharmacopeia and National Formulary*. 40th ed. 2017-02.
- [4] European Directorate for the Quality of Medicines & HealthCare. “2.2.46. Chromatographic Separation Techniques”. In: *European Pharmacopeia*. 5th ed. Strasbourg, 2005.
- [5] The Ministry of Health, Labour and Welfare. “2.01 Liquid Chromatography”. In: *The Japanese Pharmacopoeia*. 17th ed. 2016-04, pp. 41–44.
- [6] Dedecke Handelsgesellschaft mbH. *EP, USP und JP*. URL: <http://www.dedecke-gmbh.de/epuspjp.html> (visited on 2019-08-23).
- [7] European Directorate for the Quality of Medicines. *Membership and Observership for the European Pharmacopoeia*. 2017. URL: <https://www.edqm.eu/en/membership-observership> (visited on 2019-07-12).
- [8] The United States Pharmacopeial Convention. *About the Herbal Medicines Compendium*. 2013-04. URL: <https://hmc.usp.org/about/about-the-herbal-medicines-compendium> (visited on 2019-07-13).
- [9] European Directorate for the Quality of Medicines & HealthCare. *International Harmonisation*. URL: <https://www.edqm.eu/en/international-harmonisation-614.html> (visited on 2019-07-12).
- [10] Pharmacopoeial Discussion Group. “Working Procedures of the Pharmacopoeial Discussion Group (PDG)”. In: *Pharmeuropa* Vol. 22.No. 4 (2010-06), pp. 590–592.
- [11] The United States Pharmacopeial Convention. *Harmonization Status for General Chapters*. 2017-02. URL: <http://www.usp.org/harmonized-standards/pdg/general-chapters> (visited on 2019-07-13).

- [12] N. Rao, S. S. Mani Kiran, and N. L. Prasanthi. “Pharmaceutical Impurities: An Overview”. In: *indian journal of pharmaceutical education and research* 44 (2010-08), pp. 301–306.
- [13] Chris Byrne. *How are Pharmacopoeial monographs written, interpreted and changed?* URL: <https://www.raci.org.au/document/item/2056> (visited on 2019-10-27).
- [14] M. Farooq Wahab, Purnendu K. Dasgupta, Akinde F. Kadjo, and Daniel W. Armstrong. “Sampling frequency, response times and embedded signal filtration in fast, high efficiency liquid chromatography: A tutorial”. In: *Analytica Chimica Acta* 907 (2016-02), pp. 31–44. ISSN: 0003-2670.
- [15] Attila Felinger, Anikó Kilár, and Borbála Boros. “The myth of data acquisition rate”. In: *Analytica Chimica Acta* 854 (2015-01), pp. 178–182. ISSN: 0003-2670.
- [16] Daniel Stauffer. “Das Chromatogramm”. In: *Chromatogramme richtig integrieren und bewerten*. Ed. by Stavros Kromidas and Hans-Joachim Kuss. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2008-03, pp. 3–53. ISBN: 978-3-527-62222-1 978-3-527-31774-5.
- [17] John W. Dolan. “The Role of the Signal-to-Noise Ratio in Precision and Accuracy”. In: *LCGC Europe* 19.1 (2006-01), pp. 12–16.
- [18] D. A. Armbruster, M. D. Tillman, and L. M. Hubbs. “Limit of detection (LQD)/limit of quantitation (LOQ): comparison of the empirical and the statistical methods exemplified with GC-MS assays of abused drugs.” In: *Clinical Chemistry* 40.7 (1994-07), pp. 1233–1238. ISSN: 0009-9147, 1530-8561.
- [19] David A Armbruster and Terry Pry. “Limit of Blank, Limit of Detection and Limit of Quantitation”. In: *The Clinical Biochemist Reviews* 29.Suppl 1 (2008-08), S49–S52. ISSN: 0159-8090.
- [20] Fred E. Lytle and Randall K. Julian. “Automatic Processing of Chromatograms in a High-Throughput Environment”. In: *Clinical Chemistry* 62.1 (2016-01), pp. 144–153. ISSN: 1530-8561.
- [21] C. E. Shannon. “Communication in the Presence of Noise”. In: *Proceedings of the IRE* 37.1 (1949-01), pp. 10–21. ISSN: 0096-8390.
- [22] J. Efrim Boritz. “IS practitioners’ views on core concepts of information integrity”. In: *International Journal of Accounting Information Systems* 6.4 (2005-12), pp. 260–279. ISSN: 1467-0895. DOI: 10.1016/j.accinf.2005.07.001.
- [23] Office of the Federal Register and National Archives and Records Administration. *Code of Federal Regulations Title 21: Food and Drugs - Parts 1 to 99*. 2017-04. URL: <https://www.gpo.gov/fdsys/pkg/CFR-2017-title21-vol11/pdf/CFR-2017-title21-vol11.pdf> (visited on 2018-03-08).

- [24] Bitesize. *BBC - GCSE Bitesize: Verification*. 2014-07. URL: <http://www.bbc.co.uk/schools/gcsebitesize/ict/databases/3datavalidationrev2.shtml> (visited on 2018-03-23).
- [25] Agilent Technologies. *Chemical Analysis, Life Sciences, and Diagnostics | Agilent*. URL: <https://www.agilent.com/> (visited on 2019-02-12).
- [26] Thermo Fisher Scientific Inc. *Life Technologies*. URL: <https://www.thermofisher.com/de/de/home.html> (visited on 2019-02-18).
- [27] Agilent Technologies. *LICOP components for software development*. 2016-01. URL: <http://www.agilent.com/en-us/firmwareDownload?whid=56791> (visited on 2019-01-12).
- [28] Wireshark Foundation. *Download Wireshark*. 2016-07. URL: <https://www.wireshark.org/download.html> (visited on 2019-08-14).
- [29] Waters Corporation. *Waters Corporation : The Science of What's Possible : Waters*. URL: <https://www.waters.com/nextgen/de/de.html> (visited on 2019-03-14).
- [30] Waters Corporation. "ApexTrack Integration: Theory and Application". In: (2016), p. 25.
- [31] Agilent Technologies. *Evaluating System Suitability - CE, GC, LC and A/D ChemStation - Revisions: A.03.0x->A.08.0x*. URL: <https://www.agilent.com/cs/library/Support/Documents/a10424.pdf> (visited on 2019-03-08).
- [32] John Hinshaw. "How Do Your Peaks Measure Up?" In: *LC GC Europe* 26 (2013-09), pp. 575–582.
- [33] Shimadzu Corporation. *About Resolution, Part 1 : SHIMADZU*. 2019. URL: <https://www.shimadzu.com/an/hplc/support/lib/1ctalk/resol-1.html> (visited on 2019-05-14).
- [34] David Thomas Marehn, Detlef Wilhelm, Heike Pospisil, and Roberto Pizzoferrato. "Proving the Preclusion of Data Manipulation Using Parallel Data Acquisition in Chromatography". In: *Materials Science Forum* 941 (2018-12), pp. 2390–2394. DOI: <https://doi.org/10.4028/www.scientific.net/MSF.941.2390>.
- [35] David Thomas Marehn, Detlef Wilhelm, Heike Pospisil, and Roberto Pizzoferrato. "Double entry method for the verification of data a chromatography data system receives". In: *Journal of Sensors and Sensor Systems* 8.1 (2019-05), pp. 207–214. ISSN: 2194-8771. DOI: <https://doi.org/10.5194/jsss-8-207-2019>.
- [36] Agilent Technologies. *Raw Data*. Private Document. 2014.
- [37] Keshav Kumar. "Standardising the chromatographic denoising procedure". In: *Analytical Methods* 10.34 (2018-08), pp. 4189–4200. ISSN: 1759-9679. DOI: 10.1039/C8AY01606B.

- [38] Dionex Corporation. *Chromeleon - Tutorial and User Manual*. Vol. Version 6.40. 2001-11, p. 1524.
- [39] Waters Corporation. *2475 Multi λ Fluorescence Detector - Operator's Guide*. 2012.
- [40] George Lungu. *Introduction to Numerical Methods – Basic Filters #1*. 2010-10. URL: <http://www.excelunusual.com/basic-filters-1/> (visited on 2018-04-12).
- [41] Craig. *The RC Filter Transfer Function | Analog Zoo*. 2015-12. URL: <http://www.analogzoo.com/2015/12/deriving-the-rc-filter-transfer-function/> (visited on 2018-11-30).
- [42] Julius O Smith, Stanford University, Center for Computer Research in Music and Acoustics, Stanford University, and Department of Music. “Hamming Window”. In: *Spectral audio signal processing*. OCLC: 776892709. Stanford, Calif.: W3K, 2011. ISBN: 978-0-9745607-3-1.
- [43] Peter A. Gorry. “General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method”. In: *Analytical Chemistry* 62.6 (1990-03), pp. 570–573. ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac00205a007.
- [44] G. Vivó-Truyols, J. R. Torres-Lapasió, A. M. van Nederkassel, Y. Vander Heyden, and D. L. Massart. “Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection”. In: *Journal of Chromatography A. Chemical Separations and Chemometrics* 1096.1 (2005-11), pp. 133–145. ISSN: 0021-9673. DOI: 10.1016/j.chroma.2005.03.092.
- [45] Nyoman Gunantara. “A review of multi-objective optimization: Methods and its applications”. In: *Cogent Engineering* 5.1 (2018-01). Ed. by Qingsong Ai, p. 1502242. DOI: 10.1080/23311916.2018.1502242.
- [46] Paul Barrett. *Euclidean Distance - raw, normalized and double-scaled coefficients*. 2005-09. URL: <https://www.pbarrett.net/techpapers/euclid.pdf> (visited on 2019-05-03).
- [47] Agilent Technologies. *Agilent 1200 Infinity Series Diode Array Detectors - User Manual*. 2013-02.
- [48] E13 Committee. *Method E 1657: Practice for Testing Variable-Wavelength Photometric Detectors Used in Liquid Chromatography*. Tech. rep. ASTM International, 2011. DOI: 10.1520/E1657-98R11.
- [49] Paul G. Boswell, Dwight R. Stoll, Peter W. Carr, Megan L. Nagel, Mark F. Vitha, and Gary A. Mabbott. “An Advanced, Interactive, High-Performance Liquid Chromatography Simulator and Instructor Resources”. In: *Journal of chemical education* 90.2 (2013-02), pp. 198–202. ISSN: 0021-9584. DOI: 10.1021/ed300117b.

- [50] Jiří Vojta, Pavel Hanzlík, Aleš Jedlička, and Pavel Coufal. “Separation and determination of impurities in paracetamol, codeine and pitophenone in the presence of fempiverinium in combined suppository dosage form”. In: *Journal of Pharmaceutical and Biomedical Analysis* 102 (2015-01), pp. 85–92. ISSN: 0731-7085. DOI: 10.1016/j.jpba.2014.09.002.
- [51] Council of Europe, European Pharmacopoeia Commission, and European Directorate for the Quality of Medicines & Healthcare. *European pharmacopoeia*. 7th ed. OCLC: 662704167. Strasbourg: Council Of Europe : European Directorate for the Quality of Medicines and Healthcare, 2010. ISBN: 978-92-871-6700-2.
- [52] D. T. Marehn, D. Wilhelm, R. Pizzoferrato, and H. Pospisil. “Recommendations for optimal sampling and filter rates in liquid chromatography”. In: *AMA Service GmbH, Von-Münchhausen-Str. 49, 31515 Wunstorf*. 2019. DOI: 10.5162/sensoren2019/p3.11.
- [53] Tino Weinkauff and Yeara Kozlov. *Extracting and Filtering Minima and Maxima of 1D Functions*. 2013-11. URL: <https://www.csc.kth.se/~weinkauff/notes/persistence1d.html> (visited on 2018-09-06).
- [54] Serge Beucher and Christian Lantuéjoul. “Use of Watersheds in Contour Detection”. In: *In International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation* 132 (1979-01).
- [55] The MathWorks, Inc. *Quadratic programming - MATLAB quadprog*. URL: <https://www.mathworks.com/help/optim/ug/quadprog.html> (visited on 2019-03-15).
- [56] Jorge Nocedal and Stephen J. Wright. “Quadratic Programming”. In: *Numerical optimization*. 2nd ed. Springer series in operations research. OCLC: ocm68629100. New York: Springer, 2006, p. 449. ISBN: 978-0-387-30303-1.
- [57] Sybil P. Parker. “Laplace Operator”. In: *McGraw-Hill Dictionary of Scientific and Technical Terms*. McGraw-Hill Education, 2003. ISBN: 978-0-07-042313-8.
- [58] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. “Laplace Interpolation”. In: *Numerical recipes: the art of scientific computing*. 3rd ed. OCLC: ocn123285342. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. 150–153. ISBN: 978-0-521-88068-8.
- [59] C. T. Kelley. “Conjugate Gradient Iteration”. In: *Iterative methods for linear and nonlinear equations*. Frontiers in applied mathematics 16. Philadelphia: Society for Industrial and Applied Mathematics, 1995, pp. 11–32. ISBN: 978-0-89871-352-7.

- [60] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. “Conjugate Gradient Method for a Sparse System”. In: *Numerical recipes: the art of scientific computing*. 3rd ed. OCLC: ocn123285342. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. 87 –93. ISBN: 978-0-521-88068-8.
- [61] Richard H. Bartels, John C. Beatty, and Brian A. Barsky. “Parametric vs. Geometric Continuity”. In: *An introduction to splines for use in computer graphics and geometric modeling*. 3. Dr. OCLC: 258413197. Los Altos, Calif: Kaufmann, 1996. ISBN: 978-1-55860-400-1.
- [62] H. Ugail. “Introduction to Partial Differential Equations”. In: *Partial differential equations for geometric design*. OCLC: ocn733247655. London ; New York: Springer, 2011, pp. 21–30. ISBN: 978-0-85729-783-9.
- [63] Yeara Kozlov, Janick Martinez Esturo, Hans-Peter Seidel, and Tino Weinkauff. “Regularized Harmonic Surface Deformation”. In: (2014-08). eprint: 1408.3326.

List of Figures

1.1	Several terms for description and evaluation of a chromatogram	5
2.1	Concept of the simultaneous data acquisition	10
2.2	Exact experimental setup of the parallel data acquisition	13
2.3	Bypass the Licop library	15
2.4	Peak Detection via <i>Apex Track</i>	17
2.5	Influences of coupled sampling rate and filtration settings on the chromatogram	21
2.6	Overlayed plots of <i>OpenLab ChemStation</i> [®] and <i>Data Collector</i> for 1260 HPLC series system	22
2.7	Overlayed plots of <i>OpenLab ChemStation</i> [®] and <i>Data Collector</i> for 1200 HPLC series system	23
2.8	Overlayed plots of <i>Chromeleon</i> [®] and <i>Data Collector</i> for 1260 HPLC series system	24
2.9	Overlayed plots of <i>Chromeleon</i> [®] and <i>Data Collector</i> for 1200 HPLC series system	25
2.10	Overlayed plots of analog-digital-converter and "Data Collector" for 1260 HPLC series system	26
2.11	Scatter plot of analog-digital-converter and <i>Data Collector</i> for 1260 HPLC series system after time shift	27
2.12	In-Depth analysis of data storage by <i>Chromeleon</i> [®] and disabled averaging .	29
3.1	Recommendation of Sampling Rate and Signal Filtration	34
3.2	Low Pass RC Filter Circuit [41]	36
3.3	Multi objective optimization forming a Pareto Optimal Front (POF)	40
3.4	Response Time Determination	43
3.5	<i>SmooSi Maker</i> Application Screen	45
3.6	Simulated Separation of Paracetamol, Codeine, Pitophenone and Impurities .	48
3.7	Separation of Methyl Paraben and Prednison	51
4.1	Self-constructed example plot	55
4.2	Concept of the <i>Persistence 1D</i> algorithm based on the watershed transforma- tion displayed on Caffeine spectrum	56

4.3	Inverted plot of fig. 4.1 and highlighted local maxima and minima	57
4.4	Relevant maximum-minimum pairs and corresponding components	58
4.5	Filtered raw data and laplacian reconstruction	61
4.6	Comparison laplacian, biharmonic, triharmonic reconstruction	63
4.7	Comparison laplacian, biharmonic, triharmonic reconstruction using a Data Weight of 0.01	65
4.8	Persistence filtration and reconstruction depending on DW	66
4.9	Isocratic Test Standard	67
4.10	Application of Persistence Filter on small peak	68
A.1	Influences of sampling rate settings on the chromatogram	84
A.2	In-Depth analysis of data storage by Chromeleon [®] and enabled averaging . .	90

List of Tables

2.1	Chromatographic conditions for simultaneously data acquisition	14
2.2	Single run conditions for Licop checkup	16
3.1	Kind of filters related to the HPLC vendors	36
3.2	Experimental conditions for separation of Prednison and Methyl paraben . .	47
3.3	Optimization related to active substances	49
3.4	Optimization related to impurities	50
3.5	Application of recommendation in practice	52
4.1	Persistence Filtration of Isocratic Test Standard	67
4.2	Filter influences on large peak	69
4.3	Filter influences on small peak	69
A.1	Comparison of Data Collector and Ethernet data packages	85
A.2	Simulation of 21 peaks	91
A.3	Simulation of one large and two small peaks	93
A.4	Simulation of one large and one very small peak	93

Appendix

A.1 Peak Efficiency of Data Acquired by ADC

Figure A.1: Influences of sampling rate settings on the chromatogram

Both chromatograms have been acquired by the Analog-To-Digital Converter (ADC) *35900 Dual Channel Interface* manufactured by *Agilent Technologies*. The peak efficiencies (black numbers), retention times red numbers and signal-to-noise values of 10 injections are shown. The peak efficiency reduces a little bit when a lower sampling rate of 10 Hz is used whereby the retention times are nearly the same. The signal-to-noise values S/N_1 to S/N_4 on the other hand increases significantly when a lower sampling rate is used even no filter algorithm is applied to the signal by the ADC.

A.2 Data Packages Caught from the Ethernet Connection

Table A.1: Comparison of Data Collector and Ethernet data packages

The following table shows the data acquired by the *Data Collector* and the bytes caught from the Ethernet connection using the sniffer tool *WireShark*. The single data points are saved in 32 bit (4 byte) values. These values are shown in hexadecimal and decimal format here. Based on the behavior that 2097.152 counts correspond to 1 mAU the values have been converted. The values in milli absorbance units (mAU) of the *Data Collector* and caught from the Ethernet connection are completely equal.

Data Collector	Ethernet		
[mAU]	[Counts] _{HEX}	[Counts] _{DEC}	[mAU]
0.161170959472656	00 00 01 52	338	0.161170959472656
0.160694122314453	00 00 01 51	337	0.160694122314453
0.191688537597656	00 00 01 92	402	0.191688537597656
0.902652740478516	00 00 07 65	1893	0.902652740478516
8.052349090576170	00 00 41 f7	16887	8.052349090576170
35.695075988769500	00 01 24 6a	74858	35.695075988769500
90.866088867187500	00 02 e8 60	190560	90.866088867187500
161.372184753418000	00 05 29 f6	338422	161.372184753418000
223.629951477051000	00 07 27 fa	468986	223.629951477051000
263.221740722656000	00 08 6c 50	552016	263.221740722656000
277.694702148438000	00 08 e2 e0	582368	277.694702148438000
260.958671569824000	00 08 59 c6	547270	260.958671569824000
211.350440979004000	00 06 c3 62	443234	211.350440979004000
143.843650817871000	00 04 9a 5e	301662	143.843650817871000
83.312034606933500	00 02 aa 7e	174718	83.312034606933500
44.140815734863300	00 01 69 9a	92570	44.140815734863300
23.308753967285200	00 00 be f2	48882	23.308753967285200
12.988567352294900	00 00 6a 67	27239	12.988567352294900
7.875919342041020	00 00 40 85	16517	7.875919342041020
5.235671997070310	00 00 2a e4	10980	5.235671997070310
3.773212432861330	00 00 1e e9	7913	3.773212432861330
2.893447875976560	00 00 17 b4	6068	2.893447875976560
2.321243286132810	00 00 13 04	4868	2.321243286132810
1.928329467773440	00 00 0f cc	4044	1.928329467773440
1.646041870117190	00 00 0d 7c	3452	1.646041870117190
1.439094543457030	00 00 0b ca	3018	1.439094543457030
1.280784606933590	00 00 0a 7e	2686	1.280784606933590
1.157283782958980	00 00 09 7b	2427	1.157283782958980
1.062870025634770	00 00 08 b5	2229	1.062870025634770

Data Collector	Ethernet		
[mAU]	[Counts] _{HEX}	[Counts] _{DEC}	[mAU]
0.999927520751953	00 00 08 31	2097	0.999927520751953
0.968933105468750	00 00 07 f0	2032	0.968933105468750
0.961780548095703	00 00 07 e1	2017	0.961780548095703
0.974178314208984	00 00 07 fb	2043	0.974178314208984
0.996589660644531	00 00 08 2a	2090	0.996589660644531
1.021385192871090	00 00 08 5e	2142	1.021385192871090
1.039028167724610	00 00 08 83	2179	1.039028167724610
1.050472259521480	00 00 08 9b	2203	1.050472259521480
1.059055328369140	00 00 08 ad	2221	1.059055328369140
1.066207885742190	00 00 08 bc	2236	1.066207885742190
1.076698303222660	00 00 08 d2	2258	1.076698303222660
1.093864440917970	00 00 08 f6	2294	1.093864440917970
1.116275787353520	00 00 09 25	2341	1.116275787353520
1.134872436523440	00 00 09 4c	2380	1.134872436523440
1.142978668212890	00 00 09 5d	2397	1.142978668212890
1.142024993896480	00 00 09 5b	2395	1.142024993896480
1.134872436523440	00 00 09 4c	2380	1.134872436523440
1.121520996093750	00 00 09 30	2352	1.121520996093750
1.103401184082030	00 00 09 0a	2314	1.103401184082030
1.083850860595700	00 00 08 e1	2273	1.083850860595700
1.067638397216800	00 00 08 bf	2239	1.067638397216800
1.053810119628910	00 00 08 a2	2210	1.053810119628910
1.040458679199220	00 00 08 86	2182	1.040458679199220
1.025676727294920	00 00 08 67	2151	1.025676727294920
1.007080078125000	00 00 08 40	2112	1.007080078125000
0.985145568847656	00 00 08 12	2066	0.985145568847656
0.961780548095703	00 00 07 e1	2017	0.961780548095703
0.937938690185547	00 00 07 af	1967	0.937938690185547
0.912666320800781	00 00 07 7a	1914	0.912666320800781
0.884532928466797	00 00 07 3f	1855	0.884532928466797
0.856399536132813	00 00 07 04	1796	0.856399536132813
0.831127166748047	00 00 06 cf	1743	0.831127166748047
0.806808471679688	00 00 06 9c	1692	0.806808471679688
0.782489776611328	00 00 06 69	1641	0.782489776611328
0.760078430175781	00 00 06 3a	1594	0.760078430175781
0.737190246582031	00 00 06 0a	1546	0.737190246582031

Data Collector	Ethernet		
[mAU]	[Counts] _{HEX}	[Counts] _{DEC}	[mAU]
0.714302062988281	00 00 05 da	1498	0.714302062988281
0.689983367919922	00 00 05 a7	1447	0.689983367919922
0.663757324218750	00 00 05 70	1392	0.663757324218750
0.638008117675781	00 00 05 3a	1338	0.638008117675781
0.612258911132813	00 00 05 04	1284	0.612258911132813
0.591754913330078	00 00 04 d9	1241	0.591754913330078
0.576972961425781	00 00 04 ba	1210	0.576972961425781
0.566005706787109	00 00 04 a3	1187	0.566005706787109
0.554084777832031	00 00 04 8a	1162	0.554084777832031
0.539302825927734	00 00 04 6b	1131	0.539302825927734
0.524520874023438	00 00 04 4c	1100	0.524520874023438
0.510215759277344	00 00 04 2e	1070	0.510215759277344
0.494480133056641	00 00 04 0d	1037	0.494480133056641
0.479221343994141	00 00 03 ed	1005	0.479221343994141
0.467300415039063	00 00 03 d4	980	0.467300415039063
0.462055206298828	00 00 03 c9	969	0.462055206298828
0.460147857666016	00 00 03 c5	965	0.460147857666016
0.458240509033203	00 00 03 c1	961	0.458240509033203
0.454425811767578	00 00 03 b9	953	0.454425811767578
0.448703765869141	00 00 03 ad	941	0.448703765869141
0.441074371337891	00 00 03 9d	925	0.441074371337891
0.431537628173828	00 00 03 89	905	0.431537628173828
0.421524047851563	00 00 03 74	884	0.421524047851563
0.413417816162109	00 00 03 63	867	0.413417816162109
0.408649444580078	00 00 03 59	857	0.408649444580078
0.407218933105469	00 00 03 56	854	0.407218933105469
0.406265258789063	00 00 03 54	852	0.406265258789063
0.404357910156250	00 00 03 50	848	0.404357910156250
0.399589538574219	00 00 03 46	838	0.399589538574219
0.391483306884766	00 00 03 35	821	0.391483306884766
0.380992889404297	00 00 03 1f	799	0.380992889404297
0.370502471923828	00 00 03 09	777	0.370502471923828
0.363826751708984	00 00 02 fb	763	0.363826751708984
0.358581542968750	00 00 02 f0	752	0.358581542968750
0.354290008544922	00 00 02 e7	743	0.354290008544922
0.351905822753906	00 00 02 e2	738	0.351905822753906

Data Collector	Ethernet		
[mAU]	[Counts] _{HEX}	[Counts] _{DEC}	[mAU]
0.350475311279297	00 00 02 df	735	0.350475311279297
0.350475311279297	00 00 02 df	735	0.350475311279297
0.349521636962891	00 00 02 dd	733	0.349521636962891
0.350475311279297	00 00 02 df	735	0.350475311279297
0.350952148437500	00 00 02 e0	736	0.350952148437500
0.349521636962891	00 00 02 dd	733	0.349521636962891
0.344276428222656	00 00 02 d2	722	0.344276428222656
0.335693359375000	00 00 02 c0	704	0.335693359375000
0.325679779052734	00 00 02 ab	683	0.325679779052734
0.315189361572266	00 00 02 95	661	0.315189361572266
0.306129455566406	00 00 02 82	642	0.306129455566406
0.300407409667969	00 00 02 76	630	0.300407409667969
0.298500061035156	00 00 02 72	626	0.298500061035156
0.298976898193359	00 00 02 73	627	0.298976898193359
0.299930572509766	00 00 02 75	629	0.299930572509766
0.300884246826172	00 00 02 77	631	0.300884246826172
0.300884246826172	00 00 02 77	631	0.300884246826172
0.299930572509766	00 00 02 75	629	0.299930572509766
0.298500061035156	00 00 02 72	626	0.298500061035156
0.297546386718750	00 00 02 70	624	0.297546386718750
0.296115875244141	00 00 02 6d	621	0.296115875244141
0.293254852294922	00 00 02 67	615	0.293254852294922
0.288009643554688	00 00 02 5c	604	0.288009643554688
0.281810760498047	00 00 02 4f	591	0.281810760498047
0.276565551757813	00 00 02 44	580	0.276565551757813
0.272750854492188	00 00 02 3c	572	0.272750854492188
0.269889831542969	00 00 02 36	566	0.269889831542969
0.267982482910156	00 00 02 32	562	0.267982482910156
0.268936157226563	00 00 02 34	564	0.268936157226563
0.272274017333984	00 00 02 3b	571	0.272274017333984
0.275135040283203	00 00 02 41	577	0.275135040283203
0.276088714599609	00 00 02 43	579	0.276088714599609
0.275135040283203	00 00 02 41	577	0.275135040283203
0.272274017333984	00 00 02 3b	571	0.272274017333984
0.268936157226563	00 00 02 34	564	0.268936157226563
0.266075134277344	00 00 02 2e	558	0.266075134277344

Data Collector	Ethernet		
[mAU]	[Counts] _{HEX}	[Counts] _{DEC}	[mAU]
0.265598297119141	00 00 02 2d	557	0.265598297119141
0.267505645751953	00 00 02 31	561	0.267505645751953
0.270366668701172	00 00 02 37	567	0.270366668701172
0.273704528808594	00 00 02 3e	574	0.273704528808594
0.276565551757813	00 00 02 44	580	0.276565551757813
0.277996063232422	00 00 02 47	583	0.277996063232422
0.277519226074219	00 00 02 46	582	0.277519226074219
0.274658203125000	00 00 02 40	576	0.274658203125000
0.271320343017578	00 00 02 39	569	0.271320343017578
0.267028808593750	00 00 02 30	560	0.267028808593750
0.262260437011719	00 00 02 26	550	0.262260437011719
0.256538391113281	00 00 02 1a	538	0.256538391113281
0.248908996582031	00 00 02 0a	522	0.248908996582031

A.3 In-Depth analysis of Chromeleon®

Figure A.2: In-Depth analysis of data storage by Chromeleon® and enabled averaging

The data provided by 1200 series HPLC series system acquired by *Chromeleon*® and additionally caught from the Ethernet connection by *WireShark* using two different sampling rate and signal filtration configurations (a and b). The average parameter of *Chromeleon*® is enabled here. The apparent time shifts have been normalized (c and d). The still existing deviations are plotted in (e) for both sampling rates.

A.4 Chromatogram Simulation Parameters

Table A.2: Simulation of 21 peaks

Simulation parameters for chromatogram showing separation of three active substances Paracetamol, Codeine and Pitophenone and 18 Impurities. The chromatogram has a sampling rate of 40 Hz.

Parameter	Value	Parameter	Value
Peak 1		Peak 7	
Flow	1	Flow	1
Number of Moles	0.05	Number of Moles	0.023
Retention Time	120	Retention Time	540
Peak Width	0.35	Peak Width	0.35
Peak 2		Peak 8	
Flow	1	Flow	1
Number of Moles	14	Number of Moles	0.023
Retention Time	240	Retention Time	750
Peak Width	14	Peak Width	0.35
Peak 3		Peak 9	
Flow	1	Flow	1
Number of Moles	0.03	Number of Moles	0.06
Retention Time	300	Retention Time	780
Peak Width	0.35	Peak Width	0.35
Peak 4		Peak 10	
Flow	1	Flow	1
Number of Moles	0.02	Number of Moles	10
Retention Time	3600	Retention Time	1020
Peak Width	0.35	Peak Width	10
Peak 5		Peak 11	
Flow	1	Flow	1
Number of Moles	0.005	Number of Moles	0.08
Retention Time	450	Retention Time	1140
Peak Width	0.35	Peak Width	0.35
Peak 6		Peak 12	
Flow	1	Flow	1
Number of Moles	0.025	Number of Moles	0.035
Retention Time	480	Retention Time	1260
Peak Width	0.35	Peak Width	0.35

Parameter	Value	Parameter	Value
Peak 13		Peak 18	
Flow	1	Flow	1
Number of Moles	0.09	Number of Moles	0.012
Retention Time	1320	Retention Time	1680
Peak Width	0.35	Peak Width	0.35
Peak 14		Peak 19	
Flow	1	Flow	1
Number of Moles	0.01	Number of Moles	0.06
Retention Time	1360	Retention Time	1740
Peak Width	0.35	Peak Width	2
Peak 15		Peak 20	
Flow	1	Flow	1
Number of Moles	0.04	Number of Moles	10
Retention Time	1380	Retention Time	1860
Peak Width	0.35	Peak Width	10
Peak 16		Peak 21	
Flow	1	Flow	1
Number of Moles	0.035	Number of Moles	0.008
Retention Time	1560	Retention Time	1950
Peak Width	0.35	Peak Width	0.35
Peak 17		Noise	
Flow	1	Standard Deviation	0.0001
Number of Moles	0.02		
Retention Time	1650		
Peak Width	0.35		

Table A.3: Simulation of one large and two small peaks

The listed simulation parameters are used to check out the influence of the Data Weight (DW) parameter of the new *Persistence Filter*. The sampling rate of the simulated chromatogram is 100 Hz.

Parameter	Value
Peak 1	
Flow	1
Number of Moles	10
Retention Time	12
Peak Width	1.8
Peak 2	
Flow	1
Number of Moles	0.09
Retention Time	20
Peak Width	0.3
Peak 3	
Flow	1
Number of Moles	0.05
Retention Time	30
Peak Width	0.35
Noise	
Standard Deviation	0.001

Table A.4: Simulation of one large and one very small peak

The listed simulation parameters are used to check out the influence of the threshold and Data Weight (DW) parameters of the new *Persistence Filter* on peaks very next to the baseline noise. The sampling rate of the simulated chromatogram is 100 Hz.

Parameter	Value
Peak 1	
Flow	1
Number of Moles	4
Retention Time	12
Peak Width	0.35
Peak 2	
Flow	1
Number of Moles	0.02
Retention Time	30
Peak Width	0.5
Noise	
Standard Deviation	0.003