# Analysis of Plastic Recycling and Breast Cancer Prognosis using Vibrational Spectroscopy

Chuang Gao

Thesis submitted to The University of Nottingham

for the degree of Doctor of Philosophy

Sept. 2014 – Sept. 2019

"*Left to our own devices... we are all too good at picking out non-existent patterns that happen to suit our purposes.*"

– Efron and Tibshirani, 1993

Dedicated to the LGBTQ+ community

*keep reaching for that rainbow!*

# Abstract

This thesis is split into three parts.

### Part I: An Automatic Self-optimising Continuous-flow Reactor for Electrosynthesis

Assembly-line and/or continuous, steady-state strategies are widely spreading in the manufacturing industry. Electrosynthesis in an automatic continuous flow manner is drawing more and more attention to both pharmaceutical industry and research. This part describes the develop and use of an automated self-optimising continuous-flow electrochemistry reactor system. The main focus of this work is to develop an bi-language (MATLAB and LabVIEW), server-client structure automation software framework to conduct automation, control and monitoring of flow electrochemistry processes, which enables quick system setup, reconfiguration and high flexibility. Stable noisy optimisation by branch and fit (SNOBFIT) and simplex algorithm (modified simplex and super-modified simplex method) were developed and tested on simulators, which were then applied to the synthesis action of methoxylation of N-formylpyrrolidine

and electro-oxidation of 3-bromobenzyl alcohol. Searching of the optimum operation condition of a reaction in an automatic manner is a major step forward to establish convenient and straightforward use of organic electrosynthesis in routine laboratory synthesis or industrial applications.

## Part II: Applying FTIR Imaging to Address Challenges in Plastic Recycling

Plastic pollution is ubiquitous throughout the earth, and reusing/recycling of plastic has the potential to reduce the global abundance and weight of waste plastics. The main focus of this work is investigating plastic sample using Fourier transform infrared spectroscopy (FTIR), single point and imaging, for recycling purpose. An quantitative calibration of talcum concentration in talcum reinforced virgin polypropylene sample with IR peak ratio/integration was conducted, and the application of the result to analyse the talcum disperse in polypropylene matrix was reported. Micron scale FTIR imaging was conducted on the film sample. Pseudo-colour image visualising the distribution of talcum in the polypropylene matrix indicated a highly uneven distribution, a result of the reprocessing method.

FTIR imaging was applied to investigate the composite structures of 'real-world' composites sample for recycled industry plastics, including: virgin polypropylene with short milled recycled carbon fibre, virgin polypropylene, maleic anhydride grafted polypropylene with carbon fibre, acrylonitrile butadiene styrene with calcium carbonate and virgin polypropylene with poly(ethylene tereph-

thalate) on the micron scale. Imaging technique in FTIR spectroscopy not only provide micron level spatial information of the composition but also direct solution of improving the inter-facial interactions between compositions, thus improving the physical/chemical performance of the plastic products. Those pilot studies provide insights into applying FTIR imaging for plastic sample analysis.

## Part III: FTIR spectroscopy for Breast Cancer Prognosis

Breast cancer is a major cause of deaths for females worldwide. Cancer prognosis provides a patient's likely outcome based on their current standing, which can help to decide the treatment for the patient. The current golden standard prognosis index, Nottingham Prognosis Index, is a time-consuming, un-objective process to which limited confidence can be assigned because of inherent operator variability. Applying Fourier transform infrared spectroscopy (FTIR) imaging to breast cancer tissue offers a non-destructive, label free tool for cellular and extracellular breast cancer tissue studying. In this work, we evaluate the prognostic ability of FTIR spectroscopy for identifying different grades of breast cancer. Different combinations of data pre-processing, feature extraction and unsupervised learning methodologies are explored. Spectrum quality control methods are applied to correct or minimise spectral problems, including high noise level, baseline offset and outlier. A multi-stage data analysis algorithm developed can provide statistical control over the breast cancer classification process and produce a precise cancer prognosis.

# Abbreviations

Below are listed the abbreviations commonly used throughout this Thesis. Any abbreviations that are not listed are given in the text at their first occurrence.

| | |
|---|---|
| **IR** | Infrared |
| **FTIR** | Fourier transform infrared spectroscopy |
| **ATR** | attenuated total reflectance |
| **SNR** | signal-to-noise ratio |
| **FPA** | focal plane arrays |
| **H&E** | haematoxylin and eosin |
| **SVM** | support vector machine |
| **PCA** | principal components analysis |
| **NPI** | Nottingham Prognosis Index |
| **FE** | feature extraction |
| **FS** | feature selection |
| **FC** | feature construction |
| **DS** | diagnostic system |
| **GA** | genetic algorithm |
| **ANN** | artificial neural networks |

| | |
|---|---|
| **LDC** | linear discriminant classifier |
| **FPPE** | Formalin fixed, paraffin embedded |
| **QCL** | quantum cascade laser |
| **VPP** | virgin polypropylene |
| **MAPP** | maleic anhydride grafted polypropylene |
| **PET** | polyethylene terephthalate |
| **EN** | Epithelial |
| **LN** | Lymphocyte |
| **ELV** | end-of-life vehicles |
| **WEEE** | waste electrical and electronic equipment |
| **FOV** | Field Of View |
| **SEM** | scanning electron microscope |
| **CF** | carbon fibre |
| **rSMCF** | Recycled short milled carbon fibre |
| **FCM** | Fuzzy-C Mean clustering |
| **SF** | surfactant |
| **ABS** | acrylonitrile butadiene styrene |
| **SNOBFIT** | stable noisy optimisation by branch and fit |
| **MSM** | Modified Simplex Method |
| **SMS** | Super-Modified Simplex Method |
| **NMSIM** | Nelder-Mead simplex algorithm |

# Acknowledgements

Working together with Prof. Mike George for the past five years has been a fantastic journey. He's smart, but not self-righteous. He's busy, but always on the top of your research. If you see through the surface, one can be deeply touched by how much he cares and how much efforts he puts in training you into a better researcher.

Dr. Peter Summers is my second supervisor. Even though his research field is far away from mine, the amount of time we spent together to go through the smallest details of my research is tremendous. Not mentioning he spent his first research grant buying me a workstation that I desperately needed for my research.

Dr. Philip Hall was closely involved in the plastic project. It is all those fruitful discussion with him and his student Yitao lead the project to its success.

Dr. Ke Jie was my team leader on the self-optimisation reactor project. All the discussion and advices from him helped me a lot. I also thank Mr. Bin Wang, who was the collaborator on the case study of VPP + Talc project; My undergraduate supervisor Prof. Shuhong Wang, who is not directly involved into my Ph.D project, but all the training that I get from him benefited me a lot

in my Ph.D.

My best friend Michael, who introduced me into English language and culture, who helped me went through bad relationships and dark moments when I was suffering from culture shock and light depression; my best friend Kate, who taught me and still teaching me to care, to reach out and to be frank to myself and everyone around me; my friends Matt and Jorge, for all the beautiful time we spent together in NingBo and all my other friends/colleagues whom I met in my first year in computer science, second year and third year in UNNC and final year in chemistry school, I thank you all.

I thank everyone in George and Poliakoff group. Those cheerful talks with Jonny, the wonderful yoga time with Amy, the tremendous help from Katie, Charlotte C. and Charlotte P, the morning coffee everyday from Paul and Ashley, this list can go on and on ... I am so lucky to have you all.

I thank my parents. Unlike any other Chinese parents who over-control everything, they gave me freedom and trust. I'm able to choose what I want, love what I love and do what I want to do because of them.

It is the love from my family, the care from my supervisors and the support from friends and colleagues, give me the courage to continue fighting for knowledge, for better, and for the future.

# Declaration

I declare that the work presented in this Thesis is a record of my own work for the period of September 2014 to September 2018. Where any work has been carried out in collaboration, the relevant researcher or researchers have been acknowledged. This Thesis has not been accepted in partial or complete fulfilment or any other degree or personal qualification.

Chuang Gao, 3rd of January 2019

# Contents

# Part I

# AN AUTOMATIC SELF-OPTIMISING CONTINUOUS-FLOW REACTOR FOR ELECTROSYNTHESIS

CHAPTER 1

# Introduction

Green chemistry is a relatively new research field that aims to harness chemical innovation to meet environmental and economic goals [1] simultaneously. It involves 'reducing or eliminating the use or generation of hazardous substances in the design, manufacture and application of chemical products' [1]. Anastas and Warner proposed The Twelve Principle of Green Chemistry to help chemists achieve the intentional goal of sustainability:

1. **Prevention.** It is better to prevent waste than to treat or clean up waste after it is formed.

2. **Atom Economy.** Synthetic methods should be designed to maximise the incorporation of all materials used in the process into the final products.

3. **Less Hazardous Chemical Synthesis.** Whenever practicable, synthetic methodologies should be designed to use the generate substances that pose little or no toxicity to human health and the environment.

4. **Designing Safer Chemicals.** Chemical products should be designed to

preserve efficacy of the function while reducing toxicity.

5. **Safer Solvents and Auxiliaries.** The use of auxiliary substances should be made unnecessary whenever possible and, when used, innocuous.

6. **Design for Energy Efficiency.** Energy requirements of chemical processes should be recognised for their environmental and economic impacts and should be minimised. If possible, synthetic methods should be conducted at ambient temperature and pressure.

7. **Use of Renewable Feedstocks.** A raw material or feedstock should be renewable rather than depleting whenever technically and economically practicable.

8. **Reduce Derivatives.** Unnecessary derivatisation should be minimised or avoided if possible, because such steps require additional reagents and can generate waste.

9. **Catalysis.** Catalytic reagents are superior to stoichiometric reagents.

10. **Design for Degradation.** Chemical products should be designed so that at the end of their function they break down into innocuous degradation products and do not persist in the environment.

11. **Real-Time Analysis for Pollution Prevention.** Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances.

12. **Inherently Safer Chemistry for Accident Prevention.** Substances and the form of a substance used in a chemical process should be chosen to min-

imise the potential for chemical accidents, including releases, explosions, and fires.

In accordance with the Principles of Green Chemistry, the advantages of synthesis in continuous flow is a 'green' alternative to batch processes in terms of Energy Efficiency and Prevention, which have attracted the attention of chemists [2, 3]. The use of continuous flow offers several advantages ranging from controlled process conditions to high flow rates and mass throughput [4]. It enables reactions to be performed with an unprecedented level of control, affording excellent transferability between laboratory-based investigations and subsequent production scales [5]. Other benefits of reaction miniaturisation include reduced exposure to hazardous chemicals, through the use of sealed reactor units, increased atom efficiency as a result of precise reaction control and the ability to incorporate online analytics to closely monitor processes [5]. Microreactors, a new class of continuous reactors that have emerged over the past decade, provide continuous processing, controlled and efficient mixing of reagents, and ease of operation for the synthesis of, for example, quantum dots [6] and silica particles [7], as well as rapid screening of the reaction space with minimal consumption of scarce reagents, e.g., oligosaccharides [8].

A key aspect of this chapter is the optimisation of continuous flow electrochemistry synthesis using an automatic continuous-flow processes. As a new emerging technology in continuous-flow chemistry, a continuous flow system for electrochemistry synthesis is described in detail. Combining this system with self-optimising algorithms enables the researching of the optimal operating condition in an automatic manner. The generation of a software frame-

work, as well as the application of optimisation algorithms, is fundamental to producing an automated continuous-flow reactor system that can self-optimise performance of industrially relevant reactions. This work develops a server-client software structure, which enables quick system setup, reconfiguration and high flexibility of the continuous flow chemistry reactor system. Application of SNOBFIT, as well as Simplex Method optimisation algorithms on continuous-flow reactor for the purpose of reaction optimisation were conducted. We have investigated an electrosynthesis of methoxylation of N-formylpyrrolidine and oxidation of 3-bromobenzyl alcohol. The automated continuous flow reactor rig conducted the optimisation of both reactions successfully with less than 30 experiments, resulting in the best experimental conditions (based on the criterion used) for each reaction.

## 1.1 Continuous flow chemistry

In continuous flow chemistry, the reagent mixture is continually pumped through a reactor, and subjected to a series of tightly controlled conditions [9–11]. It is normally used to scale up processes and to conduct extreme pressure/temperature experiments safely [12–18]. One of the most important differences between continuous reactors and batch vessel processes is the kinetic (rates of chemical processes) profile of the reaction. In a batch vessel, the reaction conversion changes as a function of time, but in a flow reactor, conversion becomes a function of reactor length and the residence time [19]. Fluid dynamics determine the characteristics of continuous-flow equipment, the pressure loss, residence

time, heat transfer and mixing time [20]. One well-known example of continuous flow chemistry is the synthesis of the natural product pristane. Flögel et al. [21] employed the continuous flow dehydration of alkanols as a key reaction step in the synthesis of 2,6,10,14- tetramethylpentadecane. Using a microfluidic approach, the authors reported the synthesis of pristane in 80% yield from farnesol. Compared to conventional batch techniques, this synthetic route proved advantageous as it only requires a simple purification to be conducted unlike the multiple distillations previously employed.

In a flow reactor, pressure and temperature can be safely manipulated far beyond atmospheric conditions [22, 23]. Reactions done in flow are often faster than in the corresponding batch reactions, which gives improved energy, time, and space efficiency [24]. There is increasing interest in adapting continuous flow processes to minimise the amount of waste generated in chemical processes [4, 25, 26]. Detailed reviews on what kind of chemistries can be run continuously and what equipment is available to do so are thoroughly covered [27–30]. Reviews on when and why one should utilise flow for sustainable synthesis and selected examples can be found [26].

Environmental Impact Factor, what is now widely accepted as the E-factor was first introduced by Sheldon [31] to quantify the amount of waste generated per kilogram of product. A higher E-factor means more waste and greater negative environmental impact. The ideal E-factor is zero. Oil refining and bulk chemicals industries, by using continuous flow processes, have managed to achieved E-factors of <0.1, and <5 respectively [32]. In bulk chemistry, nearly all chemical processes benefit from continuous operation [4]. Tilstam and co-workers gave

a recent example of such. Indoles and phenols are important transformations often carried out on a large scale using hazardous reagents such as methyl iodide or dimethyl sulfate [26]. Dimethylcarbonate (DMC) has been recognised as a green, albeit less reactive alternative. Due to the relatively low boiling point (90 $^{o}$C) and reduced reactivity of this reagent, methylation reactions with DMC are generally slow [33]. Tilstam [34] used a flow reactor to perform phenol and N-heterocycle methylations with DMC. At 220 $^{o}$C, yields up to 97% could be obtained with reaction times as short as ten minutes. The solvent remains stable under the reaction conditions thus preventing the formation of an unwanted impurity.

Another most common metric used to evaluate the environmental impact of a process is a life cycle assessment (LCA). Here, two or more processes are compared from a holistic point of view, and the comparative impact on factors such as global warming or resource depletion is evaluated [26]. Kralisch and coworkers looked at the environmental impact of performing the epoxidation of soybean oil (produced commercially at a rate of approximately 240 000 tonnes per year) in batch and flow [35]. The energy demand per mole of the product was found to be lowest when performed at high temperatures in a flow reactor (T > 100 $^{o}$C). In a best-case scenario, the authors note that switching the existing process to a high temperature flow reaction can give approximately 11–12% reduction in global warming and human toxicity potential.

In comparison, the pharmaceutical and fine chemical industries have E-factors that can range between 5 and 100 [32]. The majority of pharmaceutical products are produced using a start and stop batch process [9–11], which is partly the

result of a highly dynamic and rapidly changing market in modern chemical synthesis [36, 37]. It has been estimated that the use of continuous flow manufacturing could result in saving of up to 40% in terms of cost savings in each drug loading/key-intermediate price scenario [38]. Further development of methods in continuous processing was identified as the most important area of research in green chemistry and engineering for the pharmaceutical industry [39].

It has been a growing development in the use of continuous flow reactors in the manufacture of fine chemicals and pharmaceuticals. In 2005, the American Chemical Society (ACS), Green Chemistry Institute (GCI) and several leading global pharmaceutical corporations developed the ACS GCI Pharmaceutical Roundtable to encourage innovation while catalysing the integration of green chemistry and green engineering into the business of drug discovery, development and production [40]. Some companies have already successfully developed continuous commercial-scale processes such as SK Chemicals, Ampac Fine Chemicals, and Phoenix Chemicals. SK-Chemicals has chosen a global approach through technology fusion including catalysis, microreaction technology, simulated moving bed, etc. [41]. An industrial example at Lonza was given by Kockmann et al. [4] to illustrate the capabilities of pharmaceutical production in microstructured continuous flow devices. Using organometallic and coupling as an example of a reaction, a large-scale pilot plant with vessels (600–1600 L in size) for a production of nearly 700 kg of isolated product was achieved.

A continuous process can perform steps/procedures in an automated manner

using advanced process control and appropriate modules (dedicated equipment/ unit operations), thus reducing labour costs [42]. There is considerable interest in the automation of continuous flow reactors to enable fast parameter (temperature, pressure, and time) adjusting and achieve optimal condition for the reaction [43].

## 1.2 Process optimisation and automation

Chemical reactions can be influenced by multiple factors, and therefore, finding the optimal operation conditions for a complex reaction can be an arduous task. Process optimisation is the practice of minimising, the cost, use, equipment size, waste, and time associated with a synthetic route. With the aim of commercialisation, this is a common goal within chemistry [38]. Reactions for the purpose of discovery, optimisation, and kinetic analysis seek only to gain information about a given set of conditions, not to form large amounts of products. Using batch synthesis, a traditional search procedure involves trying different combinations of parameters (e.g. temperature, pressure, time or concentrations), which requires a high level of labour and/or expenses. When there are more than three parameters to be adjusted, in most cases, it will become impossible to conduct an exhaustive search (evaluating each possible combination) on batch synthesis. To minimise waste generation, it is desirable to perform these transformations on as small a scale as is practical. Continuous-flow operations, and the potential online analysis inherent in microsystems for chemistry (microreactors) have enabled chemical researchers to investigate not only reaction

kinetics and mechanisms studies [44, 45], but also myriad reactions under experimental conditions not easily achieved with conventional laboratory batch equipment [46]. In flow chemistry, the conditions can be varied quickly, and the installation of analytical equipment (UV [47], IR [48], Raman [49], fluorescence[50], NMR [51], HPLC [52], and MS [53] devices) can allow for online analysis and feedback. Using $\mu$L volume microreactor, large amounts of data can be gathered with small amounts of material. These characteristics make flow chemistry an ideal platform for process optimisation.

Online chromatographic methods can be performed by fabricating microreactors capable of separation and detection or by transferring samples from the microreactor system to commercial HPLC/GC equipment. Mills and Nicole [54] used a multi-port gas-switching valve to sample heterogeneous gas-phase oxidation from six parallel packed-bed microreactors. Benson et al. [55] also used online gas chromatography for real-time evaluation of catalytic transformations of lipids to biodiesel via several different zeolites in a glass microreactor.

Exploiting the transparency of silicon to IR radiation in the range of 4,000 to 10,000 $cm^{-1}$, Floyd et al. [45] achieved in-situ FTIR spectroscopy monitoring for the hydrolysis of propionyl by loading a silicon microreactor directly into the sample compartment of an IR spectrometer. Further integration of temperature control was accomplished with this device through the fabrication of an on-chip heat exchanger. Additionally, detection by multiple internal reflection–FTIR spectroscopy in silicon microreactors has been accomplished through the use of a potassium hydroxide wet etch to create the bevelled silicon edges needed for the technique [56].

Even rough optimisation requires much control and many repeat experiments to acquire enough data [57]. One solution lies in automation, which allows statistically significant data to be obtained whilst decreasing the amount of manual control and monitoring required during an experiment, by incorporating logic and feedback control with instrumented microreactors. Benefits from automation and microreactors, including reductions in material and time, have been demonstrated for reaction screening and reaction optimisation [58–60]. Multiple pieces of equipment (e.g. pumps, reactor, temperature control unit and gas-liquid separator) are normally involved in continuous flow chemical reactions. The automation of the continuous flow chemistry process requires the communication, controlling and collaboration of different pieces of equipment. This is normally achieved through in-house developed automation software.

An automated microreactor system that can quickly profile the parameter space of a Sonogashira reaction was designed by Sugimoto and coworkers [61]. This system consisted of two HPLC pumps for fluid handling, a temperature control, a 500 $\mu$m $\times$ 1000 $\mu$m micromixer followed by a 1000 $\mu$m $\times$ 10 m residence time unit, and a fraction collector. The system performed the specified reactions and collected samples in an automated fashion, before the fractions were analysed offline by HPLC. This system was used to investigate reactions with 0.225 M bromothiophene derivative, 1.08 equivalents of p-tolyacetylene, 1 mol% Pd and 2 mol% Cu catalysts, and 2.5 equivalents of base. By manually input the desired experimental matrix, namely the set of reaction temperatures and residence times of interest, the optimal yield of 96% was obtained.

The majority of work on automated continuous flow reactors [9–12, 15–18, 43,

47, 58–60, 62] have focused on the hardware as well as the chemical reaction itself. Each of these examples relies on a distinct protocol in terms of software design. The automation software is normally well fitted to one specific rig setting, which makes it task-specific and difficult to transform. Major changes of the software are needed when the rig is changed to serve a new reaction. The creation of automatic flow chemistry systems is challenging because the system contains many components that needed to be re-configured and re-connected to perform a different reaction. Control schemes for devices in an automatic continuous flow rig that offer stable operation with fast dynamics are required for the adoption of microreactor systems in total synthesis research [46].

## 1.3 Automated self-optimising continuous flow reactor software

Combining automated continuous flow chemistry with self-optimisation algorithms can remove much of the tedium of optimisation [57]. Except for the automation software discussed in the last section, the self-optimisation algorithm built in the software is another essential component.

Walsh et al. [12] reported the development of a supercritical $CO_2$ reactor for heterogeneous acid catalysed etherification reactions. Using online gas-liquid chromatography (GLC), the reactor was able to tune the key factors of the reaction to maximise the yield of products. The same reactor was applied to a range of different reactions, including hydrogenations [13], aldol condensations[15] and methylations [16–18]. In the automatic optimisations steps of these research, a single factor/variable was evaluated at a time. Only after the

first factor had been optimised could another factor be varied. This approach is only reasonable if the variables are independent of each other, which, for most chemical reactions, is not true. It is important to develop efficient algorithms for continuous flow processes to optimise multiple factors in a series of consecutive experiments. To allow numerous parameters to be optimised simultaneously, a more robust system is needed.

Krishnadasan et al. [60] developed an automated platform capable of determining optimal operating conditions of CdSe QD synthesis of different sizes. An online spectrometer was used to analysis the emergent particles. An automated feedback loop using the Stable Noisy Optimisation by Branch and Fit (SNOBFIT) was adopted to generate the new conditions. First, two-dimensional optimisations in which the optimum was discovered by varying the flow rates of the Cd and Se precursors were performed. Each optimisation trial involved 43 automated experiments and investigated the conditions that would result in the best reaction performance for CdSe QDs with target emissions of 500, 510, 520, 530, 540, and 550 nm. A three-dimensional optimisation search in which reaction temperature was included in the set of variables was then performed. In the 3-D optimisation, 106 automated experiments to identify the optimal conditions for CdSe QDs at the various target emission wavelengths. In parallel to Krishnadasan et al., we have been developing a self-optimising reactor for reactions in supercritical carbon dioxide ($scCO_2$) [63]. We have previously reported the 2-D and 3-D optimisation of fluorescent nanoparticle synthesis using a microfluidic reactor [64]. This has been applied to the methylation of 1-pentanol in supercritical carbon dioxide using dimethyl carbonate and methanol with

$\gamma$-alumina catalyst to give 98% and 68% yield of pentyl methyl ether respectively. Those study nicely illustrates the ability to quickly acquire an abundance of reaction data with these platforms.

Using the Nelder-Mead Simplex Method (NMSIM) for optimisation, McMullen et al. [46] developed a self-optimising microreactor which allows numerous parameters to be optimised simultaneously for a Heck reaction. The reaction was then successfully scaled up 50-fold using the optimal conditions. In separate studies by the same group, Simplex, SNOBFIT and Steepest Decent Methods were successfully tested to optimise a Knoevenagel condensation reaction, demonstrating the broad operational capabilities of the system [65]. In addition, a four-dimensional optimisation using the Simplex Method was performed on an oxidation reaction.

Parrott et al. [63] demonstrated the first self-optimising supercritical $CO_2$ reactor on a larger scale (ca. 0.1–0.7 kg/day) than those used in previous studies with self-optimising reactors in conventional solvents [46, 65]. Using the super modified simplex (SMSIM) algorithm, a two-factor optimisation was conducted on both products from the dehydration of ethanol (1) over $\gamma$-aluminareaction, which show that conditions can be optimised for more than one product from the same reaction mixture.

Moore and Jensen [66] explored a number of different algorithms for the optimisation of a Paal Knorr reaction. A 232 $\mu$L chip reactor was used along with a commercially available IR spectroscopy flow cell which was calibrated to determine the concentration, and thus yield, of the desired product at the outlet of the reactor. Temperatures ranging from 30 to 130 $^{o}$C and reaction times from 2

to 30 minutes were evaluated. With the most effective algorithm, only 38 experiments were required to find the optimal conditions. The highest conversions were identified at 130 $^{o}$C and 12.35 minutes, giving a conversion of 81%.

A recent example was carried out by Poscharny et al. [67], who reported the advantage of using a capillary-type13 microfluidic photochemical reactor for Paternò-Büchi reactions and a computer-assisted self-optimisation of flow rates. Based on Modified Simplex algorithm (MSIM), an online-analysis approach was developed to obtain the optimal residence time by connecting an in-line ReactIR spectrometer and a computer-based communication interface to the aforementioned flow reactor. A comparable yield of 97% after 83 min reaction time was obtained.

Mathematically, there are different optimisation techniques, such as Hill Climbing, Stable Noisy Optimisation by Branch and Fit (SNOBFIT) and Simplex Methods. Among them, SNOBFIT and Simplex Methods (including Nelder-Mead Simplex Method and Super Modified Simplex Method) are the two most used methods in self-optimising flow chemistry literature. SNOBFIT is known as a global optimum search method while Simplex Method can only return local optima. Simplex Method, however, takes much fewer iterations to reach the optima compared with SNOBFIT algorithm. For an optimisation problem that has limited or only one optimum, Simplex Method can find that optimum in a few experiments. For optimisation cases that are known for having multiple optima, SNOBFIT has a higher possibility of finding the global optimum. Both SNOBFIT and Simplex Method are discussed, compared and applied in this work.

A deep reinforcement learning approach was adopted for chemistry reaction optimisation by Zhou et al. [68]. They developed a model they call the Deep Reaction Optimizer (DRO) to guide the interactive decision-making procedure in optimising reactions by combining state-of-the-art deep reinforcement learning with the domain knowledge of chemistry. Iteratively, the DRO interacts with chemical reactions to obtain the current experimental condition and yield, and determines the next experimental condition to attempt. With extensive experiments on simulated reactions, the DRO method outperformed several state-of-the-art blackbox optimisation algorithms by using 71% fewer steps.

Schweidtmann et al. [69] demonstrated for the first time the use of multi-objective machine learning methods for the self-optimisation of two exemplary chemical reactions, by combining the recently developed Thompson Sampling Efficient Multi-Objective (TS-EMO) algorithm with an automated continuous reaction system. The proposed setup was capable of simultaneously optimising productivity (STY) and environmental impact (E-factor) or % impurity. The four-parameter optimisations efficiently converged to dense Pareto fronts within 68 and 78 experiments respectively.

As the development of self-optimising continuous flow reactor, reconfigurability of the system is getting more and attention. To mitigate this challenge, Bédard et al. [70] described a plug-and-play, continuous chemical synthesis system with an integrated combination of hardware, software, and analytics. The system software controls the user-selected reagents and unit operations (reactors and separators), processes reaction analytics (high-performance liquid chromatography, mass spectrometry, vibrational spectroscopy), and conducts

automated optimisations. The versatility of the system was achieved by the development of a universal bay, a standardised and flexible interface that can host any type of reaction module necessary for the particular chemistry being performed. Six different modules have been developed thus far: a heated reactor (up to 120 $^{o}$C), a cooled reactor (to –20 $^{o}$C), a light-emitting diode (LED)–based photochemistry reactor, a packed-bed reactor (for solid supported reagents and catalysts, as well as passive mixing), a membrane-based liquid-liquid separator (purification via extraction), and a bypass (for reagent addition in a minimal volume, mixing, or unused bay). The examination of the substrate scope in multiple reactions (C-C and C-N cross-coupling, olefination, reductive amination, nucleophilic aromatic substitution (SNAr), photoredox catalysis) and multistep sequences afforded greater than 50 compounds in high yield.

## 1.4 Electrochemistry in microreactor

Organic electrosynthesis has a long history dating back to the mid-1800s [71, 72]. Faraday observed that ethane was generated when a solution of sodium acetate was electrolyzed, and thereby performed the first electroorganic synthetic transformation [73]. Many interesting conversions have been reported over the past decades [74–76], such as the synthesis of dimeric alkanes using organic acid carboxylates as starting materials [77]. Although electrochemistry transformations provide an attractive, atom efficient approach to the synthesis of complex molecules, the ability to scale-up such reactions is hampered by the inhomogeneities that arise within the electric field when reactors are in-

creased in size; consequently, its use in the development of drug candidates and fine chemicals is limited. [5]. Microreactors have unique properties stemming from their submillimeter feature sizes and large surface areas relative to their volumes [78]. The use of miniaturised flow cells in order to achieve the desired reaction throughput without the need to increase the overall size of the reactor has the potential to enable well characterised electrochemical systems to be employed for the manufacture of novel chemical agents [29, 79].

In an electrochemical thin-layer flow cell, comprising of a Pt cathode and glassy carbon anode in parallel (with an inter-electrode distance of 80 $\mu$m), Atobe and co-workers [80, 81] subsequently reported the synthesis of 2,5-dimethoxy-2,5-dihydrofuran *via* the reduction of the methanolic solvent stream. They introduced the concept of using 'parallel laminar flow' within narrow gap reactors to separate the anolyte and catholyte streams without a physical separator [82–84]. With a flow rate typically 0.1 $cm^3$/min, the slow flow rate of solution limits the cell productivity to a few milligrams/hour.

Yoshida and co-workers [85–87], reported a series of reactions based on 'cation flow' methodology. Employing a diaphragm flow cell, reactants were introduced in rapid sequence into a carrier stream whereby highly reactive carbocations, such as N-acyliminium ions, were generated and converted into the respective product. In general, good reaction selectivity (67-100%) was achieved, although at only moderate conversions (50-75%) and slow product formation rates (5-50 mg/h).

More recently, Ammonite cells with extended channel length have been developed [88, 89]. The 1000 $\times$ 2 mm spiral channel and 0.5 mm inter-electrode

gap is achieved by employing a spacer in-between two parallel circular plate electrodes with radius 8.5 cm, as shown in Figure 1.1. It is designed for studies relating to optimising reaction conditions and for synthesising products on a scale of 100 mg - 10 g [89]. For the methoxylation of N-formylpyrrolidine reaction, 100% conversion can be achieved in a single pass, and it is possible to achieve more than 95% reaction selectivity and $\geq$ 20 g/h product formation rate [88].

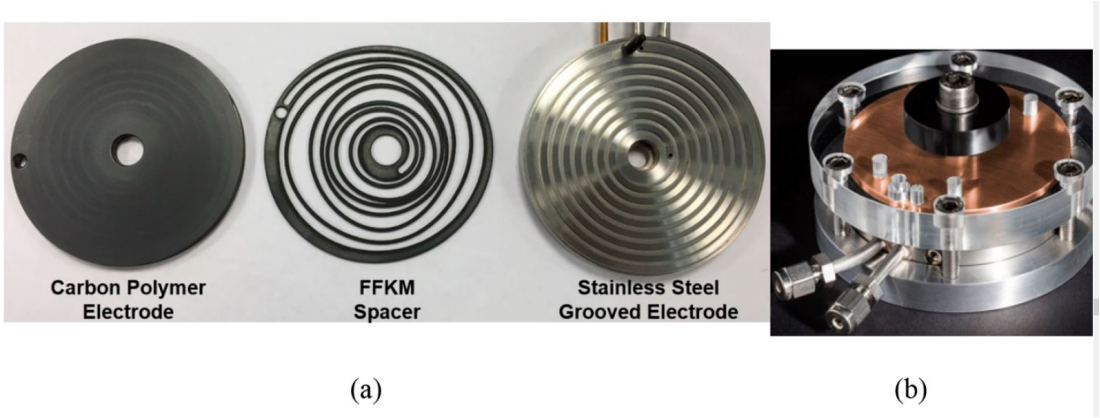



**Figure 1.1:** Ammonite cell with spiral solution channel: (a) the key components and (b) the complete cell. Reproduced from [89].

In this work, test reactions were carried out in an Ammonite electrochemical reactor. The microflow electrolysis cell for laboratory synthesis on a multigram scale has been previously developed and provided by Professor Richard C D Brown at the University of Southampton.

In addition to the reactions discussed, numerous electrochemical transformations have been investigated under continuous flow conditions, including C–C bond formations (achieved in the absence of electrolytes) [90], oxidative methoxylations [91] and cathodic reductions of allylic halides [92]. Detailed reviews

concerning organic electrode reactions [76, 93–96] and flow electrolysis cells specifically designed for the synthetic organic laboratory [71] can be found.

## 1.5 Aims

The work in this part of the thesis sets out to explain the advances made during my PhD to the automated self-optimising reactor, which can be split into two categories:

1. Rig and its automation software

   The primary aim of this research is to develop a bi-language (MATLAB and LabVIEW), server-client structure automation software framework to conduct automation, control and monitoring of the flow electrochemistry process, which enables quick system setup, reconfiguration and high flexibility of the continuous flow electrochemistry reactor system. Based on the rig used in previous studies [36, 57, 63, 64, 97–105], improvements of the hardware system of continuous flow chemistry (the electrochemical reactor, the online feedback loop, the use of a gas-liquid separator) is described in Chapter 2. This work was conducted in close collaboration with Dr. Ke Jie.

2. Optimisation strategies

   We want to explore the ways in which optimisation strategies are applied to the automated continuous flow system, and test the performances of the algorithms on varies component reactions, focusing on their chemical

consequences.

Discrete value optimisation algorithm, Stable noisy optimisation by branch and fit (SNOBFIT) was adopted and modified to fit in the self-optimisation software frame. Its efficacy and stopping criterion were investigated in a simulator. The electrosynthesis of methoxylation of N-formylpyrrolidine was used as proof of the algorithm's experimental suitableness. Continuous value optimisation algorithm, Nelder-Mead simplex algorithm (NMSIM) and Super-Modified Simplex Method (SMS) were developed and compared. NMSIM was tested on the reaction electro-oxidation of 3-bromobenzyl alcohol, and its efficacy and experimental suitableness have been investigated in comparison to the SNOBFIT algorithm.

CHAPTER 2

# Rig and automation software developing

Based on the study conducted by Ingham et al. [106], compared to perform a study/synthesis manually (batch), flow chemistry often requires more time because of the labour-intensive automation setup. The hardware of a flow chemistry system can be connected and prepared in a matter of hours or days, a smooth control and monitoring software system, however, often requires significantly more time [107]. Within the growing literature of autonomous operation of flow chemistry, the majority of the automation systems are task-specific, which is difficult for other researchers to re-invent [107]. Even a minor alteration to an automation system requires a considerable amount of effort on the software adjustments. A flexible, reliable control and automation software system is needed to reveal the full benefit of automated flow chemistry system.

Adamo et al. [3] reported a reconfigurable system for continuous flow production of pharmaceuticals. The system was divided into different modules to

enable reconfiguration to produce four different drug products within the same system. A modular software platform was used to achieve a flexible plug-and-play approach for the system. The further development of the system was reported by Bédard et al. [70]. The versatility of the system was achieved by the development of a universal bay, a standardised and flexible interface that can host any type of reaction module necessary for the particular chemistry being performed.

LabVIEW is a visual programming language popularly used in flow chemistry experiments [66, 108–111]. MATLAB has excellence in complex computational tasks. To combine the advantages of both systems, researchers place the MATLAB code inside the LabVIEW program [65, 66, 109, 111]. Cherkasov et al. [107] proposed an open-access OpenFlowChem platform based on LabVIEW for flow chemistry process automation, control and monitoring. Containing three major layers: (i) device monitors, (ii) a system module, and (iii) optional external safety devices, the platform is optimised for quick system setup, reconfiguration and high flexibility.

A LabVIEW plus MATLAB approach is adopted, as this can provide the lowest implementation costs as well as the best performance (responding time, reliability and computational resource cost). In this chapter, we propose a server-client software structure based on both MATLAB and LabVIEW programming for controlling and monitoring flow electrochemistry. Firstly, an introduction of the experiment setting for a 3-D optimisation reaction is given. Then, based on this rig setting, the software setting is explained in detail. Finally, a discussion section is provided to further explain the flexibility of the software frame.

This work is in close collaboration with Dr. Ke Jie, who is responsible for the hardware connection as well as the LabVIEW programming.

## 2.1 An example hardware connection

In this section, the electrosynthesis of methoxylation of N-formylpyrrolidine (Scheme 3.1, on page 44) was used to provide an example of the rig connection to the automated continuous-flow reactor system. The reaction is often used as an example reaction to test the performance of a flow chemistry system [112].

The use of online gas-liquid chromatography (GLC) following the progress of reactions in continuous flow has been reported [36, 64, 97, 100, 113]. In this work several improvements are made to the continuous system to incorporate an electrochemical microreactor (Figure 1.1, on page 19). Figure 2.1 shows a diagrammatic representation of the stages involved in experimentation using the automated self-optimising reactor. A typical process of automated optimisation is given as following. After the online analysis equipment returns an object function value (yield, space-time yield, E-factor, *etc.*), using a built-in optimisation algorithm, the controller predicts the next parameter combination that can potentially provide better results, and set the continuous flow reactor accordingly. The process repeats itself until the best performance is reached. In this thesis, the term 'rig' will be used to refer collectively to the experimental apparatus.

A schematic of the automated continuous-flow reactor system used in this investigation is shown in Figure 2.2a. The functional unit consists of two pumps,

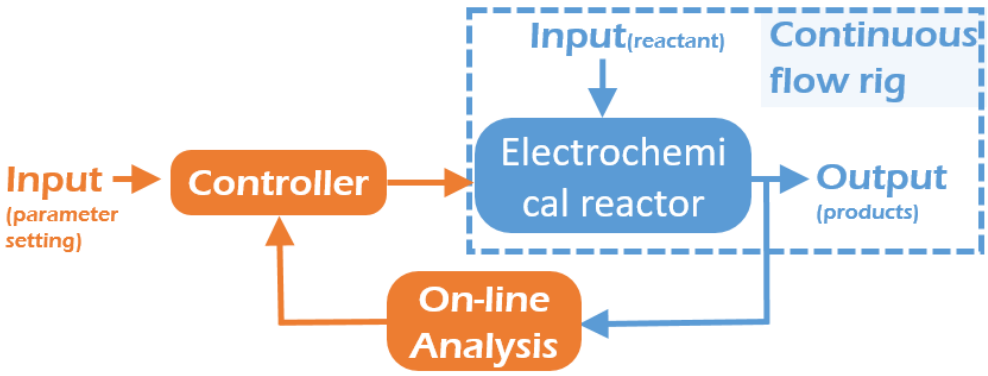
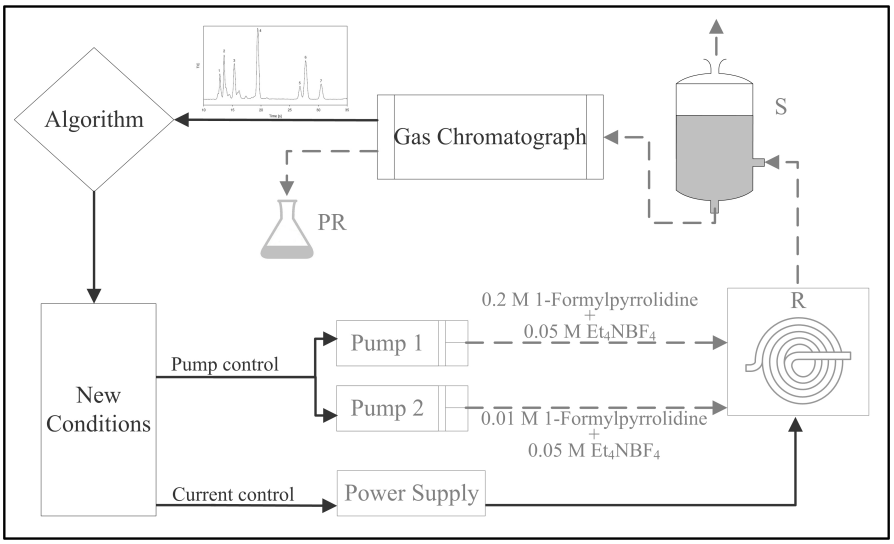
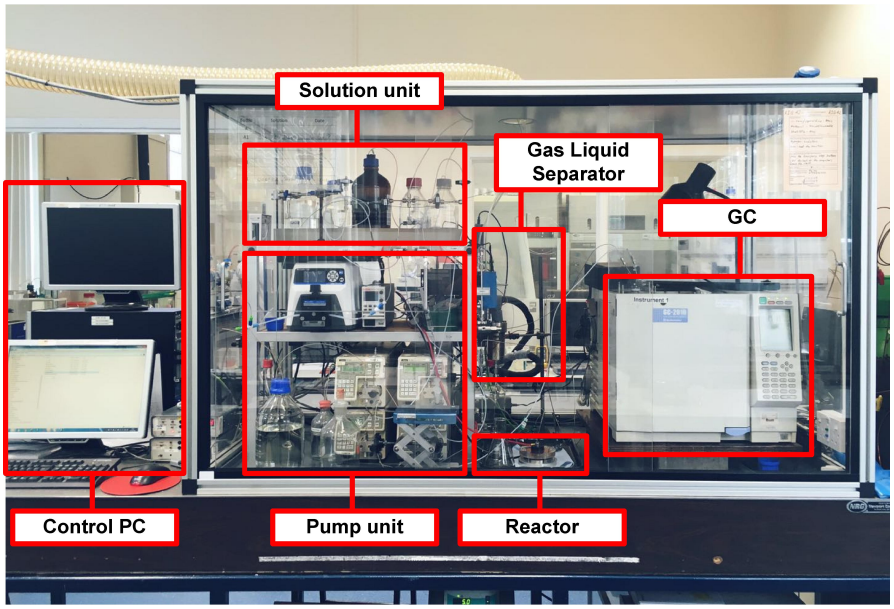



**Figure 2.1:** Schematic diagram of the continuous flow, self-optimising reactor. The blue section is a continuous flow electrochemical reactor. With constance reactant flow in, it can output the reaction result continuously into the online analysis equipment. The orange part is the automatic self-optimisation software system. Based on the on-line analysis result, the controller decides the next set of parameter setting (current, flow rate *etc.*), and applies those parameters on the electrochemical reactor. A feed-back control loop is achieved (indicated by the arrows) by the hardware and software system together to conduct chemistry reaction self-optimisation.

one reactor and one gas-liquid separator. The solvent reactant streams were pumped into the spiral micro-flow cell reactor, R, and then, to a gas-liquid separator, S. The liquid phase passes through a gas chromatography (GC) sampling loop and into to the product reservoir. Automatic sampling of the liquid phases product is achieved by GC when a new analysis is initialised. Electrolyses of the reactor was controlled with a programmable DC power supply (type 2260B-30-36, Tektronix). Communication between the central computer and the pump were accomplished with RS232 serial port connections. Online monitoring of the reaction was performed using a GC-2010 (Shimadzu) connected directly after the separator. Automation was achieved using in-house built software, which is explained in detail in Section 2.2.




**(a)**




**(b)**

**Figure 2.2:** Continuous flow, self-optimising reactor. (a) Schematic diagram of the self-optimising reactor, where the grey dash line represents reactant flow in the function unit: reactor (R), separator (S) and product reservoir (PR); and the black solid line shows the information flow between the optimising unit. (b) Experiment setting of the self-optimising reactor. Controlled by a PC, the pump unit pumps reactants into the reactor. A gas-liquid separator is used to remove the gas product, before the result solution is sent to the GC for analysis.

In this setting, the flow rate of both pump 1 and 2, as well as the current being supplied to the reaction cell are the three independent variables. Different combinations of those three parameters can be tested in an automatic, continuous flow chemistry way. Distinct GC spectrum analysis algorithms can be applied to achieve different optimisation targets (yield, E-factor etc.). Figure 2.2b is the experiment setting of the self-optimising continuous flow electrochemistry rig, with the main functioning unit labelled. The system contains six major units: (i) Control PC, fully self-sufficient LabVIEW and MATLAB programming unit, to handle all interactions with individual instruments and the self-optimisation process; (ii) Solution unit, to provide the reactant solution for electrosynthesis; (iii) Pump unit, pumps used for flow rate control, to handle input of reactant solution, output of result solution, as well as liquid level control of gas-liquid separator; (iv) Gas-Liquid Separator, to separate the gas-phase product form the result solution; (v) Electrochemistry microreactor; (vi) GC analysis equipment, to analysis the result solution for optimisation targets (yield, E-factor etc.).

## 2.2 Software developing

### 2.2.1 Requirements for flow chemistry automation

The automation of chemical synthesis is currently expanding, and this is driven by the availability of digital labware [2]. The automation software for self-optimising continuous flow electrochemistry rig not only required to establish high standard control and monitoring on multiple in-house built or commercialised lab equipments, but also needed to be safe, affordable, flexible and

accessible to synthetic chemists. Summarising the previous work on automation [12, 13, 15–18], we considered the following requirements to conduct flow chemistry in an automation way.

- Operational safety. The software should have the ability to monitor crucial experimental parameters, as well as terminate an ongoing experiment based on the feedback of those parameters. This function should have the highest priority to over-write all other instructions given. The decision-making section of this function should be carefully considered. The experiments keeps being interrupted if the entry level of the function is low (too sensitive). Failure of detecting the unsafe situation occurs when the entry level of the function is high (too insensitive).
- High reliability and stability. Robust feedback loops should be installed to conduct instrument control and communication, as software-hardware communication error can occur. For instance, a thermal meter reading function should always be called right after a temperature setting function is initialised to make sure the correct temperature is set.
- Automated pause, stop and restart function. The software should have the ability to stop or pause the experiment when the user instructs. Also, it should have the ability to restart the experiment from where it stopped. This function requires the software to record all the details of the on-going experiment and reset them to the system when re-start function is initialled by the user. This is most important for optimisation purpose as it can be expensive and time consuming to repeat the previous experiments.

– Low entry level for system alteration. Previous attempts of building software for continuous flow chemistry have shown that when one instrument is removed, the whole programming has to be carefully modified and checked. This 'enclosed' system is expensive because of the additional time and resources needed to rebuild the system. This can be avoided by the server-client software structure proposed.

– User interface. A clear and well-designed user interface should be provided to provide convenience to any user of the software.

Based on these requirements, in-house software was built for automation, control and monitoring. The software system can be divided into five main parts: user interface, Rig controller server, device driver client, optimisation client, and spectrum analysis unit. The former three are based in LabVIEW (v17.1, National Instruments), while the latter two are based in MATLAB 2017b (The MathWorks Inc., Natick, MA). The novelty of the software lies with the server-client structure software frame to make the continuous-flow synthesis system a compact, reconfigurable platform.

### 2.2.2 Software frame

As shown in Figure 2.3, we recommend a two PC system for the flow chemistry automation. Separating the commercial software of the analysis equipment (e.g. GC or IR) with the in-house build rig control system can provide extra robustness and reliability. The computer for analysis equipment is normally customised for controlling specific instrumentation. It can breakdown

easily and not up to date and it is quite costly and time-consuming to replace them. Introducing another computer into the self-optimisation reactor system, we separate the analysis instrument into a closed environment. Less system fall and high-performance stability can be expected. It also lowers the entry level for analysis equipment alteration.

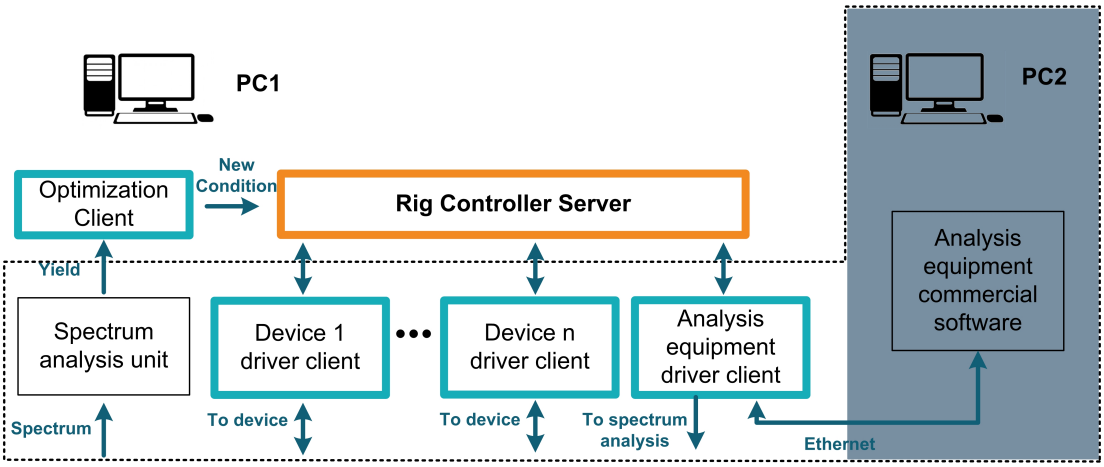



**Figure 2.3:** Software communication architecture. The system is based on a server-client structure. Distributing the system on two PCs makes the replacement of the analysis equipment/method easier. The proposed framework has five function modules: (i) drivers, (ii) rig controller, (iii) spectrum processing, (iv) optimisation algorithm, (v) user interface.

The flexibility and the transferability of the system are achieved by the server-client software structure. Rig controller server is in the middle of the automated software system. The server is consisted of all the essential functions: safety measurements, data logging, experiment stopping function, etc. Those are the functions that need no change when alternating the system. In Figure 2.3, the region highlighted by the dash line is the changeable unit, which is called a client module in our system. A new device can be easily added to the system

by adding a drive client to that device. The analysis equipment can be easily changed by changing the 'Analysis equipment drive client' and the 'Spectrum analysis unit'. A client is normally corresponding to a different type of device in the rig. It is in charge of all the automation, control and monitoring of that device. Different devices of the same type (e.g. pump 1 and 2 in Figure 2.2) can share the same client.

The optimisation function is treated as a client to the system. This enables other running methods to be used in the flow chemistry system. In optimisation methods, the next condition is only generated based upon the analysis result of the last condition. We created the other two running methods:'List' and 'Ramp'. In 'List' method, a list of the condition is provided by the user, which will be tested by the system one-by-one in an automatic manner. In 'Ramp' method, for each independent variable, a low boundary, upper boundary, and step size will be provided by the user. The method client will generate all the possible condition combinations and sent them to the rig system to be tested one-by-one. The 'List' and 'Ramp' methods proved to be useful way to determine the search boundary for self-optimisation function.

For the rig setting in Section 2.1, to conduct self-optimisation experiment, a pump client, power supply client, GC client, a gas-liquid separator client and the optimisation client are connected to the server. In a typical run, the server talks with the optimisation client to get the 'Condition' and pass the equipment parameters to the driver client to set up the experiment conditions (flow rate of pump 1, 2 and output current of power supply). After the reaction reaches steady state under the new setting conditions (10 minutes waiting), a sample

of the reactor output-flow loop is collected by the analysis equipment (GC in this case) to conducted automatic acquisition. After acquisition, the result is analysed by the spectrum analysis unit to obtain the evaluation parameter (normally yield of the target product). Based on the evaluation parameter, a 'New Condition' is generated by the optimisation client. The procedure is repeated using the new condition.

## 2.3 Conclusions on automation software developing

Using this bi-language based, server-client structured software, a continuous flow chemistry platform with minimal efforts and maximum flexibility has been achieved. The platform was demonstrated with a two parameters reaction. Both 'Optimisation' and 'Ramp' running methods were tested. The results, which is reported in Section 3.1.2 on page 44 indicate that the system can produce accurate, automatic continuous flow chemistry. Through small changes to the client (modified within hours), a three parameters optimisation experiment was conducted, which could be run (see Section 3.2.2 on page 67). This reveals the great potential of the new software platform.

CHAPTER 3

# Optimisation strategies

Parameters can be discrete or continuous values in an engineering optimisation problem. The major difference between continuous and discrete value problem is the size of the search domain. For a continuous value problem, the optimisation algorithm needs to find the best performance parameters from the infinite number of combinations. While in a discrete value problem, the search domain has limited possible combinations in which an exhaustive search can be applied. The step length is another major difference between these two types of problems. For an integer parameter problem, the step length is the integral multiple of the parameter resolution, which itself is also a discrete value. For a continuous problem, the step size can be any value decided either by the user or by the algorithm itself.

For chemistry optimisation problems, it is not always necessary to use the continuous value in optimisation. Temperature, for example, is a continuous value parameter. In some chemistry experiments, the temperature 37.150 $^{o}$C or 37.151 $^{o}$C can make little difference. Furthermore, it is difficult to control the temper-

ature at an accuracy level of 0.001 $^{o}$C. It is reasonable to fix the temperature resolution as 1 $^{o}$C or 0.5 $^{o}$C and transfer this continuous value parameter into a discrete value parameter for optimisation analysis. SNOBFIT is a discrete value optimisation algorithm, while Simplex is an optimisation algorithm works on continuous value. The applying of those two algorithms on the automated flow chemistry setting has fundamental differences in terms of the resolution. The optimisation resolution for SMOBFIT is defined by the user. Meanwhile the Simplex optimisation resolution is based on equipment accuracy grade as well as stopping criterion. This is due to the nature of the optimisation algorithm, and is further explained in the section of each algorithm.

In this chapter, both SNOBFIT and Simplex methods are compared and modified to fit in the automation software. To save time and resource, the performances of these algorithms were first tested on a simulator (return an object function value of a given condition based on mathematical equations) before real reaction optimisation is performed on the flow chemistry rig.

## 3.1 Stable noisy optimisation by branch and fit (SNOBFIT)

Stable noisy optimisation by branch and fit is the combination of a branching strategy and a sequential quadratic model. The algorithm proceeds by partitioning the search region into sub boxes, each containing exactly one point where the function has been evaluated [114]. Local minimum search is conducted by fitted quadratic models formed around data points in each box. Branching strategy is used to enhance the chance of finding the global min-

imum. The algorithm will be terminated when stopping criteria is met (e.g. not improved for four attempts). Numerical experience with further tests suggests that SNOBFIT should be used primarily with problems of dimension less than 10 [114].

The pseudo code of calling SNOBFIT optimisation algorithm is given as following:

- Generate $npoint$ random starting points, $npoint = dimensionality + i$, in which $i$ is a user define integer;
- Evaluate all the starting positions and store the best performance one as best;
- Repeat
  - Call SNOBFIT to generate new $nreq$ points, $nreq = dimensionality + j$, in which $j$ is a user define integer;
  - evaluate the new position;
  - if a value among new position is better than the best, store that position as the best;
- Stop if the user-defined termination criterion is met (e.g. not improved for 4 goes)

The SOBFIT function used in this section is an open-source MATLAB package provided by W. Huyer and A. Neumaier. Detailed instructions of using the function are given in the web page [115]. The paper describes the usage of the package, and the method implemented [114]. The MINQ bound-constrained quadratic programming package [116] used in the SNOBFIT package is MINQ5

based on MATLAB on its fifth version. The MATLAB SNOBFIT package in Reference [115] can only work with the right MINQ5 package available.

### 3.1.1 Performance on simulator

Almost every optimisation algorithm provides adjustable parameters to better fit the algorithm to a specific problem. The performance of the algorithm is highly related to the parameter settings given. Prior-knowledge of the optimisation problem/landscape is required, multiple testing of the parameter setting is necessary, before the best performance of the algorithm can be achieved. For a continuous flow chemistry rig, it is time-consuming and economically expensive to conduct optimisation algorithm parameter adjusting on real experiments. Simulators, which is a mathematical function that returns an object function value on a given condition setting, are developed and used in this section to test the performance of the optimisation method. Testing on a simulator enables the tuning of the optimisation method without conducting a real experiment.

#### Stopping criterion

Huyer and Neumaier recommended that a natural stopping test would be to quit exploration if, for a number of consecutive calls to SNOBFIT, no new point of better performance is generated [114]. The stopping criterion defined for SNOBFIT in the self-optimising software frame is as following:

*Stop the program if either one of the following is reached:*

1. *The best point did not get updated after $Numof ConCall$ times of calling SNOBFIT;*

2. *The difference between the latest and the last best point is less than I percent compared with the last one;*

3. *The limit of the number of calling object function, ncall, is reached;*

With this stopping criterion setting, there are two input parameters that need to be decided by the user: $Numof ConCall$ and $ncall$. For the SNOBFIT algorithm itself, there are another four parameters that need to be decided. In the following section, the step size of the input variables were decided before different combinations of the setting parameters were tested on the simulator.

**Step size**

SNOBFIT is an optimisation algorithm for discrete values. Users are required to define the step size for each parameter before the start of the algorithm. A big step size will lead to a smaller search domain size but have the potential problem of missing optima. A small step size can lead to a detailed search of the landscape, but have a high requirement on the number of calling the object function.

In the 2-D simulator testing below, variable 1 corresponds to the flow rate of the feeding pump, while variable 2 is the stoichiometry of current supplied to the reactor. The pump used to feed the reactants is JASCO PU-980 HPLC PUMP, which has flow rate range 1 $\mu$l/min to 10.0 ml/min with 1 $\mu$l/min step. The power supply is a programmable DC power supplies 2260B-30-36

by KEITHLEY, which has 1 mA programming resolution (by PC Remote Control Mode). Considering the resolution of the equipment, both variables were using a 0.05 step size. The step size defined by the user is only applied to points generated by SNOBFIT. The random starting points can be any number within the variables' bond range.

#### Finding the best performance parameters using simulator

An example of parameters setting 2D SNOBFIT optimisation [60] is given in Table 3.1. The search range of variable 1 and 2 were decided based on the previous study of the methoxylation of N-formylpyrrolidine reaction [112].

The mathematical equation of simulator in this section is given in Function 3.1.1. The function is based on the Goldstein-Price function, which is a commonly used 2D function for optimisation testing [117]. The function can generate a complex landscape for parameter tuning. To maintain a smooth transfer between the simulator and real experimental results, a 0.01 adjust factor was applied to the function to shift the output value into a 0-1 range. A contour plot of the function's landscape is given in Figure 3.1. Within the searching range, there are two optima: one local optimum [1.2, 0.8] and one global optimum [2.4, 0.6]. The global maximum is on the edge of the boundary to test the algorithm's performance on a constrained problem.

$$f = 0.01 \cdot \sqrt{(1 + (x_1 + x_2 + 1)^2 \cdot (19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2))} \cdot \sqrt{(30 + (2x_1 - 3x_2)^2 (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2))} \quad (3.1.1)$$

**Table 3.1:** 2D SNOBFIT parameters setting

| Name | Value | Explaination |
|---|---|---|
| NumofConCall | 4 | Limit on the consecutive call of SNOBFIT without better results |
| I | 0.001 | Termination criterion on the difference between the latest and the last best point |
| ncall | 40 | Limit on the number of object function call |
| npoint | *dimension*+4 | Number of random starting points to be generated |
| nreq | *dimension*+4 | Number of points to be generated in each SNOBFIT call |
| p | 0.3 | Probability of generating a point of class 4 |
| Variable 1 | 1-3 (step size 0.05) | Search domain of variable 1 |
| Variable 2 | 0.6-1.4 (step size 0.05) | Search domain of variable 2 |

* *dimension* is the dimension size of the problem, in the 2-D case, $dimension = 2$

* Class 1: best prediction; Class 2: putative local minimizer; Class 3: alternative good point; Class 4: explore empty region; Class 5: fill up. [114]

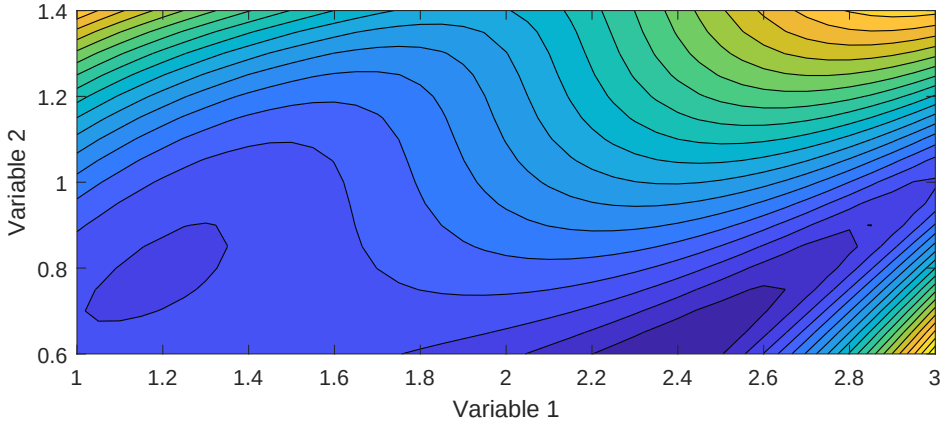



**Figure 3.1:** Contour plot of modified Goldstein-Price function. Within the selected searching range, one local optimum [1.2, 0.8] and one global optimum [2.4, 0.6] can be found. The global maximum is on the edge of the boundary, which helps to test the algorithm's performance on the boundary edge.

Using the simulator and the setting parameter described in Table 3.1, the optimisation process of finding the lowest object function value within the given variable range is conducted. Figure 3.2 is the illustration of the optimisation process. The point marked as 'start' is the best performance point between the six random starting points. Start from there, the best point was improved for three times on the experiment number 9, 16 and 19 respectively. Another 4 calling of SNOBFIT was made, and the optimisation process was ended after experiment 48, fitting the stopping criterion of '*The best point did not get updated after four times of calling SNOBFIT.*'

The performance of the algorithm is relevant to the starting point of the algorithm. As SNOBFIT randomly generates the starting points, to get rid of the influence of the starting point, the simulation was repeated for 100 times on each setting, and the average performance was used to evaluate the setting. Of

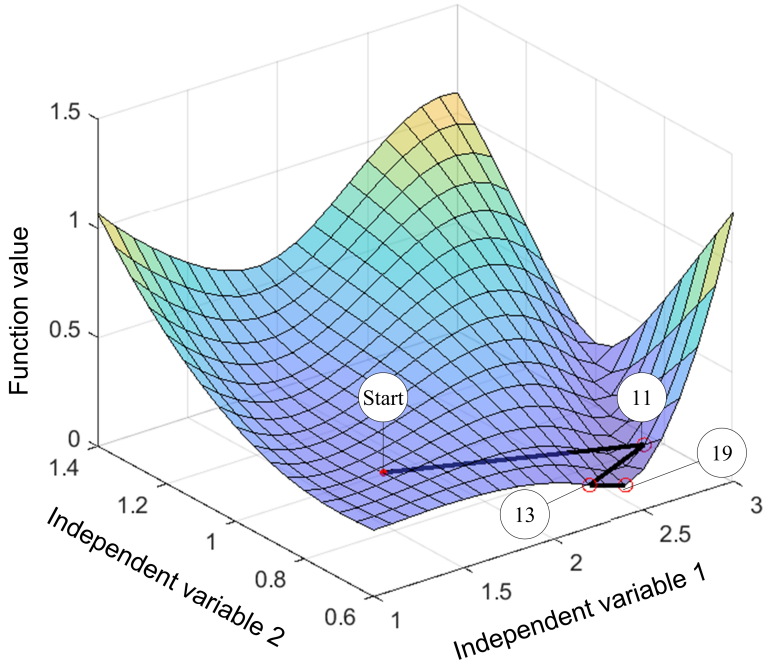



**Figure 3.2:** Simulation landscape and optimisation. With x- and y-axis representing the independent variable 1 and 2, and z-axis representing the function value, the landscape of the simulation function in that range can be plotted. Each updating of the best point is marked as red colour circle on the surface. The total number of experiments at each updating is indicated in the black circle. To tracking the improvements, a line is plotted between the current and the former best points.

the 100 times of simulation on setting parameter in Table 3.1, 52% of the simulation found the global optimum successfully. 74% of simulations stopped based on the stopping criterion '*The limit of the number of calling object function, ncall, is reached.*' This result indicates, the majority of the optimisation processes were forced to stop before the global optimum was found. This indicates that *ncall* value is too small.

The simulation was repeated for 100 times after changing the $ncall = 50$. Of the 100 times of simulation, 69% of the simulation find the global minimum suc-

cessfully. 26% of the simulation find the local minimum which is 11.14% larger than the global minimum. 80% of the simulation stopped based on the stopping criterion *'The best point did not get updated after four times of calling SNOBFIT.'* 20% of the simulation hit the maximum number of calling object function. The average times of experiments needed before the ending of optimisation is 48.24. The minimum and the maximum value are 36 and 54 respectively.

Changing the number of random starting points to be generated and the number of points to be generated in each call of SNOBFIT to $npoint = dimension + 1$ and $nreq = dimension + 1$, the 100 simulations results show that 67% of the simulation found the global minimum, 24% found the next to the best local minimum, with average times of calling object function 31.5. The minimum and the maximum value are 6 and 54 respectively. With a lower demanding of calling the object function, $npoint = dimension + 1$ and $nreq = dimension + 1$ is adopted.

With the setting $ncall = 50$, $npoint = dimension + 1$ and $nreq = dimension + 1$, 96% of the simulation stopped based on the stopping criterion 1, 2% stopped based on stopping criterion 2 and 2% stopped based on stopping criterion 3.

Attempts of increasing the percentage of finding the global optimum were tried by changing the $NumofConCall$ value into 6. The result shows a 71% of finding global minimum with a sharp rising of average times of calling object function to 39.5. With an only 4% increasing of the possibility of finding global optimum, under this combination of stopping criterion parameters setting, the optimisation process is much more time consuming and economic expensive (on average more number of experiments was needed) compared with the previous

one. This change was not adopted.

For the given simulator landscape (Equation 3.1.1), the test results show that the parameter setting (Table 3.2) can provide the best balance between the number of experiment call (average 31.5) and the possibility of finding global optimum (67%). Due to the landscape differences between the simulator and the real experiment, the performance of the SNOBFIT algorithm can be different. For most chemistry reactions, the target landscape is much less complicated compared to the simulator used in this section. The chance of finding the global optimum is higher in a real reaction optimisation than it is for the simulator. Also, the potential number of calling the experiments could be lower. Using the same parameter setting in Table 3.2, 2D SNOBFIT was tested on real-experiment data (reported in Figure 3.4 on page 48) based simulator, which has a much simpler landscape. The results show that the algorithm has 98% possibility of finding global optimum using average experiment call of 24.7. This proved the utility of testing the optimisation performance on a complex landscape simulator. The parameter settings obtained from the simulation (Table 3.2) were used to supervise the parameter setting in the real experiment conducted in the next section.

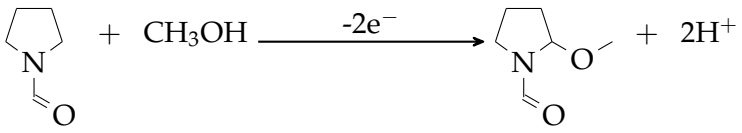
**Table 3.2:** Best performed stopping criterion setting

| **Name** | NumofConCall | I | ncall | npoint | nreq | p |
|---|---|---|---|---|---|---|
| **Value** | 4 | 0.001 | 50 | *dimension*+1 | *dimension*+1 | 0.3 |

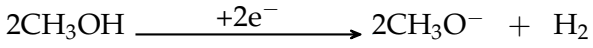
### 3.1.2 Performance on self-optimising system

The electrosynthesis of methoxylation of N-formylpyrrolidine (Scheme 3.1) was used to test the performance of the system, for comparison with previous microflow cell reactions [88, 118, 119]. The reaction was selected since it was known to give good selectivity and yield in parallel plate cells under convenient conditions [112, 120]. The cathode reaction is the reduction of MeOH to hydrogen and methoxide, so the overall chemical change is formally a dehydrogenative coupling [88].

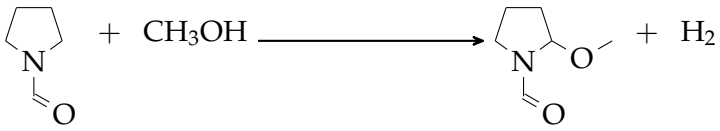
**Anode:**




**Cathode:**




**Overall:**




**Scheme 3.1:** The methoxylation of N-formylpyrrolidine.

The test reactions were carried out in the Ammonite electrochemical reactor. The $CH_3OH$ solvent reactant stream (0.1 M 1-Formylpyrrolidine + 0.05 M $Et_4NBF_4$) was pumped into the spiral micro-flow cell reactor, and then, to a gas-liquid separator.

For the SNOBFIT algorithm, two factors (flow rate and current ratio) were chosen as the independent variables. The GC peak area in percentage was used

as the objective function, which the SNOBFIT algorithm tried to maximise. The general form of the landscape function is given in Equation 3.1.2. Table 3.3 is the name of the two variables (serve as input of the automated continuous flow chemistry rig) and the response obtained from the GC measurements (output obtained from the automated continuous flow chemistry rig), together with the lower and upper bounds of each factor, which is defined based on previous research on the reaction [112].

$$\frac{A_{product}}{A_{product} + A_{reactant}} = f(FlowRate, I_{applied}/I_{theoretical}) \tag{3.1.2}$$

**Table 3.3:** Factors and response of the landscape function obtained from the reaction of the methoxylation of n-formyl pyrrolidine

| | Name | Low bound | Upper bound | Starting point |
|---|---|---|---|---|
| Factor 1 | Flow Rate (ml/min) | 1.0 | 3.0 | 2 |
| Factor 2 | $I_{applied}/I_{theoretical}$ | 0.6 | 1.4 | 1 |
| Response | $A_{product}/(A_{product}+A_{reactant})$ | - | - | - |

$A_{product}$ and $A_{reactant}$ are the peak areas of the product and the reactant, measurement by GC. $I_{applied}$ is the electrical current applied to the Ammonite electrochemical reactor. $I_{theoretical}$ represents the electrical current required to achieve 100% yield according to Faraday's laws of electrolysis, assuming that no side reactions occur. The equation for $I_{theoretical}$ calculation is given as:

$$I_{theoretical} = \frac{xnF}{t} \tag{3.1.3}$$

where $x$ is the number of moles of reactant, $n$ is the number of electrons in-

volved, $F$ is the Faraday constant and $t$ the time for the reactant solution to pass through the cell (determined by the solution flow rate of the cell) [89].

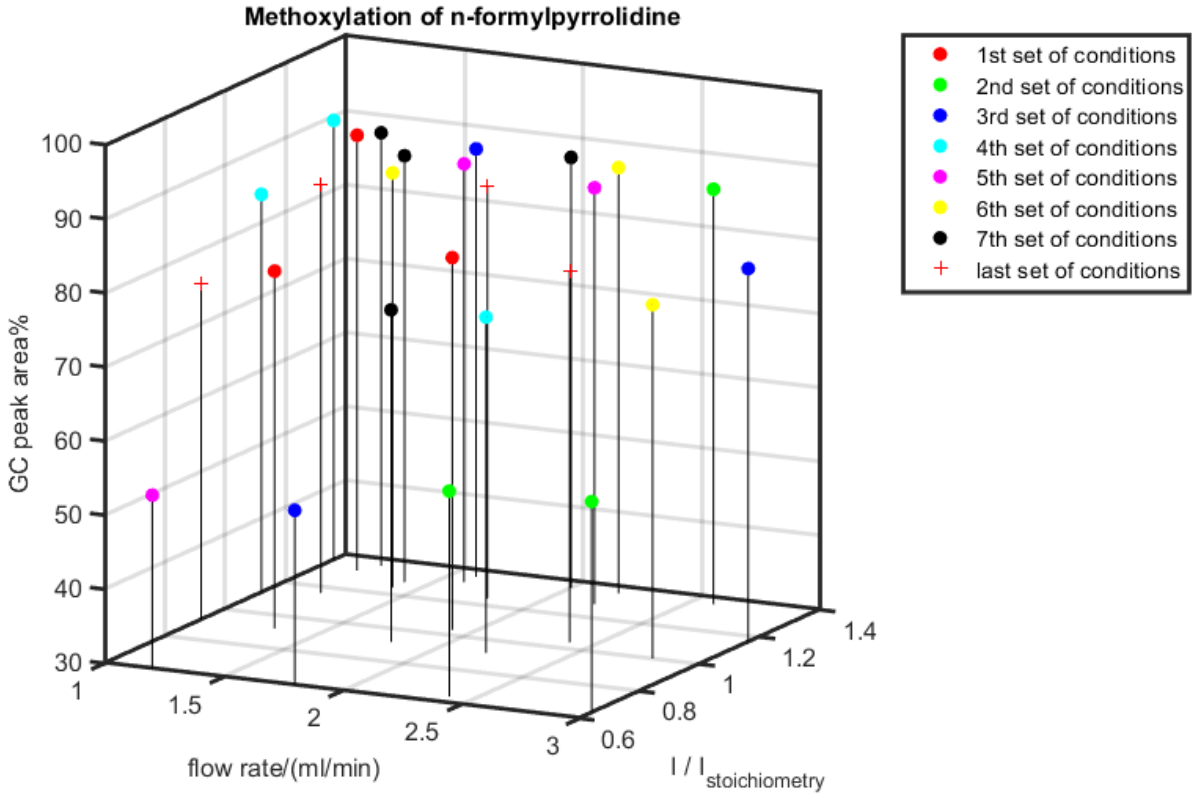



**Figure 3.3:** Experimental conditions chosen by the SNOBFIT algorithm. The two factors are the flow rate and the current ratio. (•) is the first set of experimental conditions chosen by the SNOBFIT algorithm. The other solid circles represent the 2nd to the 7th set of experimental conditions. (+) is the 8th (last) set of experimental conditions chosen by the SNOBFIT algorithm. The reaction conditions are as follows. C(N-formylpyrrolidine) = 0.1 mol/L; C($Et_4NBF_4$) = 0.05 mol/L.

With 25 hours of running, the automated continuous flow reactor rig conducted the optimisation of methoxylation of N-formylpyrrolidine successfully. The SNOBFIT algorithm ran 26 reactions, resulting in the best experimental conditions at 1.0 ml/min flow rate and 1.36 current ratio. The best GC peak area in percentage (object function value) obtained is 89%. The experiment conditions

for 26 reactions and the GC peak area in percentage measured by the on-line GC under the corresponding reaction conditions are shown in Figures 3.3. The experimental optimum was identified at the edge of the initial optimisation space, which approved the optimisation algorithm's ability to deal with the cliff edge scenario. The experiments conducted in this section is for the purpose of testing the self-optimisation rig performance, not to find the real optima of the reaction. One may consider executing a further optimisation in a larger chemical space for the latter purpose.

This result is in good agreement with the previous study on the methoxylation of N-formylpyrrolidine [88, 118, 119]. In terms of SNOBFIT algorithm, the total test performed is 26, and the simulator predicted an average of 31.5 calling of the object function. Taking the landscape difference between the reaction and the Goldstein-Price function, this is within our expectation.

Further testing of the 'Ramp' and 'List' method was carried out on this 2D rig setting, both of which turned out to be successful. For the 'Ramp' method, the first factor is the flow rate in the range of 1 to 3 ml/min, with the step size of 1; and the second factor is the ratio of the applied current to the theoretical current. The range for the second factor is 0.6 to 1.4, with the step size of 0.1. Again, the response measured for each reaction (object function value) is the GC peak area in percentage. All 27 combinations of the variables were tested, and the result is given in Figure 3.4.

The 'Ramp' test results provide a rough sketching of the object function value landscape (used as a real experiment data based simulator in Section 3.1.1). The product increases as the variable 1 (flow rate) getting lower and variable 2 (cur-

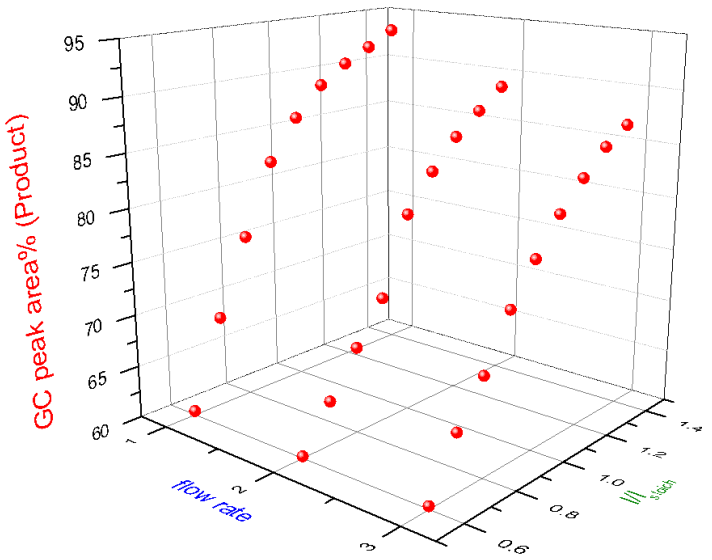



**Figure 3.4:** The GC peak area in percentage for product measured as a function of the flow rate and the current ratio. The reaction conditions are as follows. C(N-formylpyrrolidine) = 0.1 mol/L; C($Et_4NBF_4$) = 0.05 mol/L.

rent ratio) getting higher. This is in agreements with the previous electrochemical knowledge based prediction [88, 118, 119]. The best object function value obtained from the 'Ramp' test result is 93%, using 1.0 ml/min for the flow rate, 1.4 for the current ratio, which is in good agreement with the best condition obtained from the 'SNOBFIT' test method.

Successfully running SNOBFIT on methoxylation of N-formylpyrrolidine reaction optimisation indicated, not only the automation software developed can achieve automation, control and monitoring of the flow chemistry rig, but also the optimisation algorithm was well fitted to this 2D chemistry optimisation problem. The running of 'List' and 'Ramp' method proved to be rather useful pilot study to obtain a basic understanding of the object function value land-

scape and the reaction itself. Those running methods can be rather helpful in discovering unexpected and original synthetic reactions.

Electrosynthesis can allow the selective production of organic compounds under relatively mild experimental conditions and without the use of toxic/hazardous reagents [121, 122]. In order for microflow electrolysis cells to make their full contribution to routine laboratory organic synthesis, they must be capable of carrying out reactions with good selectivity and high conversion at a high rate of conversion [119]. Successful electrosyntheses are only to be expected under certain conditions. Caution is needed in generalising conclusions of one reaction to other electrosyntheses. The objective of this work was to test the self-optimising continuous flow rig developed to use in routine microflow synthesis using a model reaction, the methoxylation of N-formylpyrrolidine. This is the first self-optimisation study on organic electrosynthesis. With complete a electrosynthesis loop in less than 30 minutes, the rig is ideal for exploring the influence of reaction conditions on selectivity and yield. The searching of the optimum operation condition of a certain reaction in an automatic manner is a major step forward to establish convenient and straightforward use of organic electrosynthesis in routine laboratory synthesis or industrial applications.

## 3.2 Simplex Method

This section focuses on the development of two continuous value optimisation algorithms (Modified Simplex Method and Super Modified Simplex Method) for the self-optimisation of an electrochemical reactor. As with the SNOBFIT

algorithm above (see Section 3.1.1), a simulator is first used to test the performance of the two simplex algorithms. Following this, an optimisation experiment on the electro-oxidation of 3-bromobenzyl alcohol was carried out using the Modified Simplex Method and the parameters determined from the simulator test.

**The Basic Simplex Method (BSM)** was proposed by Spendley et al. [123]. Using $N$ (number of problem dimensions) $+1$ vertices, which is called simplex, to cover the landscape, the method can find the local optima with a relatively low number of iterations (calling the object function). Taking 2-D problem as an example, a triangle simplex shown in Figure 3.5. Assuming a maximum optima problem, the two peak points in Figure 3.5 (marked as '**x**') are the two optima containing in the landscape. The full line red triangle represents an original simplex, and the dash line triangle is the subsequent new simplex. The reflected position (the highest point in the dash line triangle, marked as '**o**') of the worst response vertex (the lowest point in the full line triangle, marked as '$*$') will be the new condition. A reflection that can improve the performance will be accepted, which means the reflected simplex (dash line triangle) will be accepted as the reflected position is higher than the worst vertex. This process repeats itself until the stopping criterion is reached.

**Modified Simplex Method (MSM)**, which is also popularly known as the Nelder-Mead simplex algorithm, was developed by Nelder and Mead [124] from the BSM method and is widely used in analytical chemistry. By allowing the simplex to change size, the modified method can adapt quickly to the shape of the responses surface, which leads to quicker searching for the local minimum.

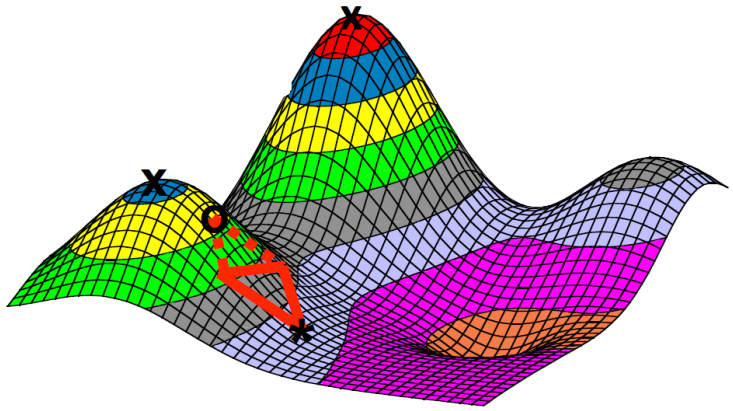


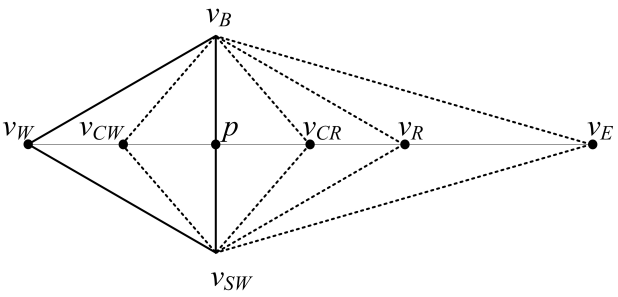
**Figure 3.5:** Demonstration of solving a 2-D problem using a triangle simplex. A reflection resulting in an uphill step (improvement) will be accepted. The full line red triangle is the original simplex, and the dash line triangle is representing the new simplex accepted as the next move. The two peak points (local optima) is marked as '**x**'. The reflected position is marked as '**o**'. The worst response vertex in the original simplex is marked as '$*$'.

A series of rules are used to define the movements of the simplex on the landscape surface. Using a 2-D simplex as an example, the spacial relationship of $v_W$, $v_R$, $v_E$, $v_{CR}$, and $v_{CW}$ are shown in Figure 3.6a. For each step, the worst response vertex $v_W$ in the current simplex can be replaced by a reflection vertex $v_R$, expansion vertex $v_E$, contraction towards the reflection vertex $v_{CR}$, and contraction towards the worst vertex $v_{CW}$. The rules listed below are used in the algorithm to decide which vertex to choose. By choosing a different vertex in different situations, the algorithm modifies the simplex size accordingly.

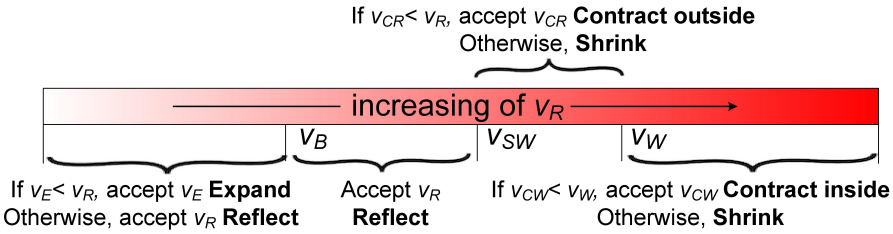
If $v_B \leq v_R < v_{SW}$, accept $v_R$ and terminate this iteration. **Reflect**

If $v_R < v_B$, calculate the expansion point $v_S$

- If $v_S < v_R$, accept $v_S$ and terminate the iteration. **Expand**




**(a)** Modified simplex for two factors. The algorithm reflects the worst vertex $v_W$ to reflection vertex $v_R$, expansion vertex $v_E$, contraction towards the reflection vertex $v_{CR}$, and contraction towards the worst vertex $v_{CW}$. $v_B$ and $v_{SW}$ is the current best and second-worst vertex respectively. $p$ is the centroid. $v_{CR}$ is the middle point between $v_R$ and $p$, while $v_{CW}$ is the middle point between $v_W$ and $p$




**(b)** Rules for selecting the next vertex. Taking $v_R$'s value as a spectrum, the lower the value the better the performance of vertex $v_R$. In the lower range of the spectrum, reflection vertex $v_R$ or expansion vertex $v_E$ is more likely to be accepted. In the higher range of the spectrum, the algorithm is likely to accept contraction towards the worst vertex $v_{CW}$ or contraction towards the reflection vertex $v_{CR}$ as the replacement of the $v_W$.

**Figure 3.6:** Modified simplex method.

- Otherwise, accept $v_R$ and terminate the iteration. **Reflect**

If $v_R \geq v_{SW}$, perform a contraction between $p$ and the better of $v_W$ and $v_R$:

- If $v_R$ is better than $v_W$, accept $v_{CR}$ and terminate the iteration. **Contract outside**. Otherwise, **Shrink**.
- If $v_R \geq v_W$, accept $v_{CW}$ and terminate the iteration. **Contract inside**. Otherwise, **Shrink**.

**Shrink**: Order the points in the simplex from the lowest function value $v_{x1}$=$v_B$ to the highest $v_{x(n+1)}$=$v_W$. Let $x(i)$ denote the list of points in the current simplex, $i = 1, ..., n+1$. $x_{shrink}(i) = x(1) + (x(i) - x(1))/2$. Use The $x_{shrink}$ as the simplex at the next iteration.

in which, $v_{SW}$ is the second-worst vertex in the current simplex.

The rule of selection is based on the performance of reflected point R. Taking $v_R$'s value as a spectrum, for different $v_R$'s value ranges, different vertex is selected. As the algorithm is designed to find the minimum in the landscape, the lower the $v_R$'s value, the better the performance of vertex $v_R$. The selecting rule of 2-D simplex is visualised, and is given in Figure 3.6b.

The MATLAB function 'fminsearch' uses MSM to find the local minimum of a multi-variable problem. It attempts to return a vector that is a local minimiser of the mathematical function near the starting vector [125].

By definition, 'fminsearch' is an unconstrained method. In the original MSM paper, boundary violations were handled by assigning a poor response to the vertex whose location violates the boundary constraint [124]. The new vertex is subsequently located between the centroid ($p$ in Figure 3.6a) and original point.

This approach is improved by applying simple mathematical transformations to convert a bound constrained problem into an unconstrained problem.

John D'Errico developed a function called 'fminsearchbnd', which is based on 'fminsearch', aiming to apply bounds to variables by transferring the problem mathematically [126]. The bounds are inclusive inequalities, which admit the boundary values, but any values outside. Variables that have both lower and upper bounds are transferred using a sine function. Variables that have only lower or upper bound are transferred use a quadratic transformation.

In the 'fminsearchbnd' program, the constrained parameter is first normalised by the following Equation 3.2.1 to adjust the value into $-1 \leqslant x \leqslant 1$ range.

$$x_{normalised} = \frac{2 \cdot (x - LB)}{UB - LB} - 1 \tag{3.2.1}$$

in which, $x_{normalised}$ is the normalised value of $x$, LB is the lower bound and UB is the upper bound.

Then arcsine is applied to transfer the constrained variable to the unconstrained domain:

$$x_{unconstrained} = 2\pi + sin^{-1}(x_{normalised}) \tag{3.2.2}$$

in which, the $x_{normalised}$ space is $-1 \leqslant x \leqslant 1$, while the result $x_{unconstrained}$ can be any value.

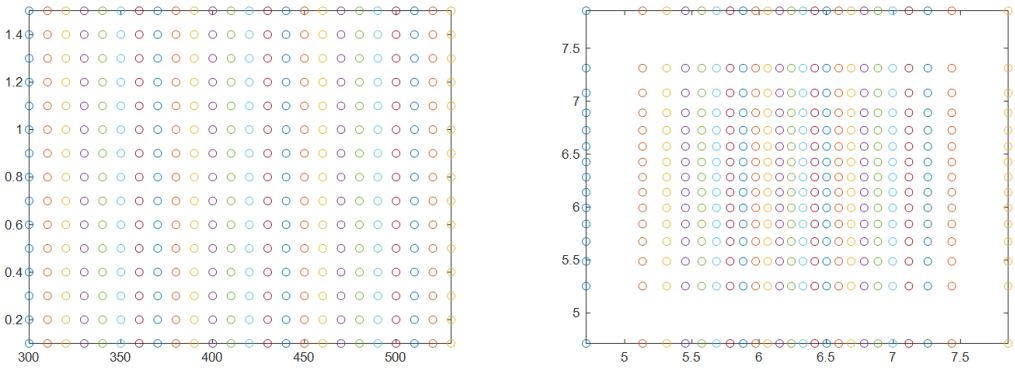
The original simplex method is input value depending. The value of the parameter influences the performance and the stopping of the algorithm. It is im-

portant to keep a balanced search domain between different parameters, which is difficult to achieve in a real engineering application. Normalising all parameters' value into $-1 \leqslant x \leqslant 1$ range is a necessary step to keep a balance between input dimensions.

One thing worth noting is that the arcsine transformation is a non-linear transformation, which makes the transferring of the simplex from the unconstrained domain into the constrained domain non-linear. Figure 3.7 shows how an original search space transforms into the new search space, which will be supplied to the unconstrained optimisation algorithm using a problem with two factors as an example. The first factor is temperature with the lower and upper bounds of 313 K and 523 K, respectively. The second factor is the flow rate in the range between 0.1 and 1.5 ml/min. In order to visualise how the original factor is transformed into the new search space, the grid size of 10 K and 0.1 ml/min have been selected and plotted.

One of the popularly used stopping criteria for simplex is to measure the size of the simplex. When the size of the simplex is smaller than the user-defined criterion (in the unconstrained domain), the optimisation process will stop. For the same simplex, different locations in the unconstrained domain will cause the size of the simplex varies in the constrained domain.

The 'fminsearchbnd' function uses the stopping criterion of 'fminsearch' directly, which means the stopping criterion based on the size of the simplex is based on the simplex size in the unconstrained domain. When transferring the stopping simplex from the unconstrained domain into a constrained domain, in a highly close to linear region, the stopping simplex will be linearly trans-




**(a)** **(b)**

**Figure 3.7:** (a) Original search space and (b) New search space after transformation using the D'Errico approach. The x-, and y- axis represent factor 1 (temperature) and factor 2 (flow rate), respectively. The axis label is ignored as the factor lost its physical meaning after transfer. Note that the scales for x-, and y- axis are very different in (a). The non-linear transfer can be observed by the position changing of dots in the space. The evenly distributed dots in the original space have non-even inter-distance in the transferred space.

formed, however, in a highly non-linear region, the equivalent constrained domain simplex is getting smaller as the location of the simplex getting closer to bounds. For an optimum close to the bounds, the optimisation algorithm may spend more iterations than necessary to reach it, as the arcsine transforms the simplex into a bigger simplex in the unconstrained domain which is further away from the stopping criterion. This problem can be solved by defining the stopping criterion in the constrained domain different from the unconstrained domain. This is achieved by converting simplex variables into their original domains before calculating the size of the simplex in 'fminsearch' function. The modified version of MSM function call is named as 'fminseachmodified' and attached in the Appendix.

The optimisation problem for continuous flow reactor rig is a constrained problem. John D'Errico's function, 'fminsearchbnd', is used to apply MSM to our system. Instead of calling 'fminsearch', 'fminsearchmodified' is called in 'fminsearchbnd' to control the size of stopping simplex in the constrained domain.

**Super-Modified Simplex Method (SMS)** is one of the many improved version of modified simplex algorithm. It can match the shape of the response surface more closely and quickly by fitting second-order polynomial [127] or Gaussian [128] to the points $v_W$, $v_R$ and centroid $p$. Compared with the basic simplex method, super-modified simplex takes fewer iterations to reach the optimum [64]. The disadvantage of the SMS method is well-recognised [129, 130]. Experiments at the centroid conditions have to be carried out before the next vertex can be decided. Although fewer iterations are needed using SMS method, the number of calling of the object functions is not necessary less compared with

the MSM method.

A detailed tutorial of using SMS is given by Morgan et al. [129]. Based on this tutorial, combining the built-in function 'fminsearch', SMS MATLAB code was developed. The function's name is 'SMS'. The code of the function is attached in Appendix.

Just like MSM, SMS is an unconstrained method defined by its nature. D'Errico's method is applied to transfer the constrained problem to an unconstrained one. The function's name is 'SMSbnd'. The code of the function is attached in Appendix.

**Simplex method with simulated annealing** is the combination of basic simplex method or super-modified simplex method with simulated annealing. By accepting downhill steps according to the simulated annealing algorithm, the change of finding the global maximum is improved [131].

Both MSM and SMS tend to find the local optima rather than the global optimum. Simulated annealing could help to overcome the problem. However, due to the nature of the simulated annealing algorithm, a large number of experiments (normally more than 100) are needed before the method reaches convergence. Each of methoxylation of N-formylpyrrolidine experiments takes more than 40 minutes, and so it is beyond practice to do so many experiments in one go.

A multi-start method could be adopted to replace simulated annealing for global optimum searching. Multiple random starting points should be tried before a global optimum is accepted. This is discussed in detail in the simulator testing

section.

### 3.2.1 Performance on simulator

Similar to SNOBFIT, Simplex methods provide setting parameters to users to fit the algorithm's purpose. Those parameters are discussed and tested on simulators in this section. Unlike SNOBFIT, which is a global optimum searching algorithm, the Simplex method can only return local optima. The number of calling object function (funcCount) alone is used to evaluate the performance.

Stopping criterion setting, as well as the starting point is required from the user in a Simplex based optimisations. Those two are discussed using the Goldstein-Price function (Equation 3.1.1) simulator. Rosenbrock function simulator was used to explore Simplex method's ability on higher dimension problems.

#### Stopping criterion

Following the 'fminsearch' function and literature, the stopping criterion used for both MSM and SMS are:

*Stop the program if either one of the following is reached:*

1. *the maximum coordinate-difference between the current best point, and the other points in the simplex is less than or equal to StepTolerance (TolX).* ***AND*** *the corresponding difference in function values is less than or equal to FunctionTolerance (TolFun).*

2. *The limit of the number of experiments (MaxFunEvals) is reached;*

3. *The limit of the number of iterations (MaxIter) is reached;*

In programming, iteration means the repetition of the optimisation process. For simplex methods, each iteration generates a new simplex. Multiple calls of the object function are needed during each iteration. MaxFunEvals is the stopping maximum value of the number of calling the object function (funcCount), while MaxIter is the maximum value of simplex used in the optimisation.

With the stopping criterion setting, there are four setting parameters that need to be decided by the user: MaxFunEvals, MaxIter, TolX and TolFun. The definition of those parameters can be found in the stopping criterion above.

In the MATLAB optimisation toolbox, the default value for both TolX and TolFun is $10^{-4}$. The default value for MaxFunEvals and MaxIter is $200 \times dimension$. More information can be found in MATLAB page on 'Tolerances and Stopping Criteria' [132].

For self-optimising reactors, each experiment loop takes over 40 minutes, including setting the parameters, waiting for steady state, collecting the sample, analysing the sample and feeding back the result to the optimisation algorithm. MaxFunEvals=100 and MaxIter=40 are used to make the optimisation process practical, which is decided based on the maximum allowance of the continuous running time of the rig.

As shown in Figure 3.7 on page 56, the transformation of the constrained problem into an unconstrained domain problem (in order to apply Simplex method, which can only process unconstrained domain problem) is non-linear. This non-linear transformation causes the algorithm to perform differently in the

same setting. To avoid this, the original 'fminsearch' is modified as 'fminseach-modified' to calculate the TolX and TolFun in the constrained domain. Using TolX=0.01, TolFun=0.01, MaxFunEvals=100 and MaxIter=40 as the stopping criterion, [2.6 1.2] as the starting point, Equation 3.1.1 as response surface simulator, performance of 'fminsearch' and 'fminseachmodified' were tested. The results of the tests are given in Table 3.4. The result indicating that for an optimum that lies on the edge of bound, fewer iterations (experiment call) are needed when defining the 'TolX' and 'TolFun' in the constrained domain.

**Table 3.4:** Stopping criterion defined in unconstrained and constrained domain

| | Output | | | |
|---|---|---|---|---|
| | x | fval | iteration | funcCount |
| fminsearch | [2.3841, 0.6000] | 0.1388 | 36 | 69 |
| fminsearch modified | [2.3832, 0.6001] | 0.1389 | 25 | 49 |

Different TolX, TolFun, MaxFunEvals and MaxIter value setting were tested on 'fminsearchmodified' to find the best setting combination. The results are summarized in Table 3.5 and Table 3.6. The range of variable 1 is 1-3, and 2 is 0.6-1.4. The starting point is fixed to [1.5 0.7] to eliminate the influence of starting point.

From the setting parameter tuning experiment results, we can see that, when the tolerance values (in stopping criterion 1) is too small, the simplex size tends to get very small before the stopping criterion is satisfied. For a real reaction optimisation, the flow rate (variable 1) and current (variable 2) cannot achieve

**Table 3.5:** Stopping criterion for MSM, with the best performance combination setting highlighted.

| | Stopping Criterion | | | | Output | | | |
|---|---|---|---|---|---|---|---|---|
| | TolX | TolFun | MaxFun Evals | MaxIter | x | fval | itera tions | func Count |
| 1 | default | default | default | default | [2.3841, 0.6000] | 0.1388 | 44 | 99 |
| 2 | 0.01 | 0.01 | 100 | 40 | [2.3841, 0.6000] | 0.1388 | 36 | 69 |
| **3** | **0.05** | **0.05** | **100** | **40** | **[2.3842, 0.6000]** | **0.1388** | **29** | **56** |
| 4 | 0.10 | 0.10 | 100 | 40 | [2.9180, 0.9374] | 0.2623 | 8 | 16 |

this level of control accuracy. However, when the tolerance value is too big, the final best performance point is not exactly on the optimum point due to the final simplex size is too big to local the vertex accurately on the optimum. The higher the tolerance value, the further away the final output from the optimum point. Much fewer iteration and function calling are required before the stopping criterion is reached. Balancing between the funcCount and the accuracy level of the final optima, the best combination of stopping criterion parameter setting is highlighted in Table 3.5 and 3.6.

Another interesting pattern indicated by the testing results is that compared with MSM, SMS is more sensitive to changing of the TolX and TolFun value. Under the same parameter settings, the final output from SMS is further away from the optima compared with MSM algorithm. This is due to the difference of simplex defining rules in the two algorithms.

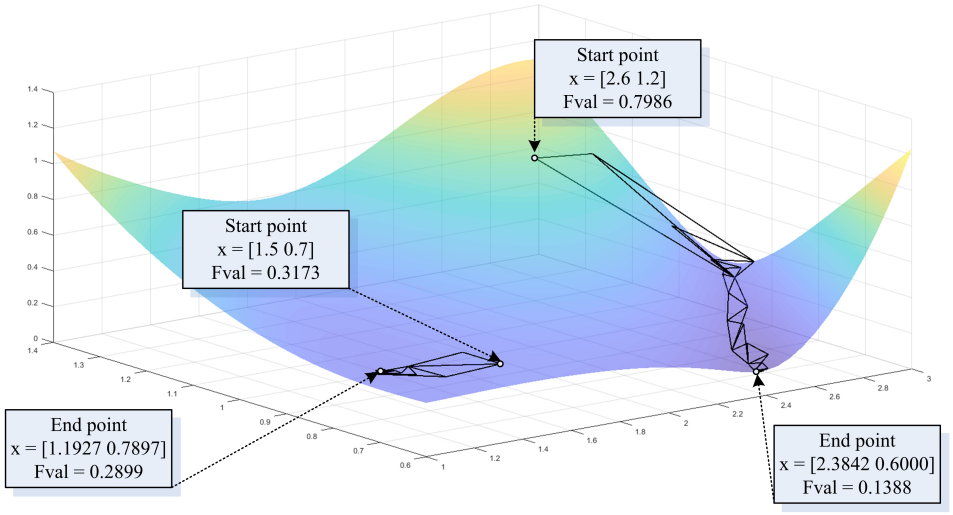
**Table 3.6:** Stopping criterion for SMS, with the best performance combination setting highlighted.

| | Stopping Criterion | | | | Output | | | |
|---|---|---|---|---|---|---|---|---|
| | TolX | TolFun | MaxFun Evals | MaxIter | x | fval | itera tions | func Count |
| 1 | default | default | default | default | [2.3770, 0.6000] | 0.1390 | 134 | 402 |
| 2 | 0.01 | 0.01 | 100 | 40 | [2.3765, 0.6000] | 0.1390 | 34 | 102 |
| **3** | **0.05** | **0.05** | **100** | **40** | **[2.3690, 0.6023]** | **0.1404** | **26** | **78** |
| 4 | 0.10 | 0.10 | 100 | 40 | [2.7875, 0.8916] | 0.2440 | 8 | 24 |

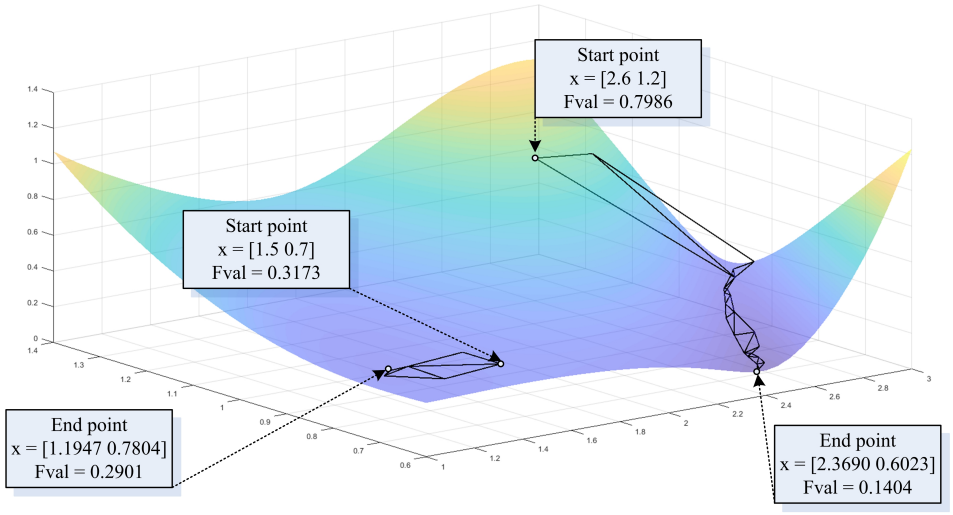
**Starting point**

A starting point is needed to initialise the simplex algorithms. Both MSM and SMS will return the local optimum found near the starting point. The choosing of the starting point can influence the algorithm's final output. It is necessary to try multiple starting points before claiming the best performance point in the constrained range is found. Within the range of 1-3 (variable 1) and 0.6-1.4 (variable 2), the 2-D simulator Function 3.1.1 has two local optima: Fval=0.1388, x=[2.3841 0.6000] and Fval=0.2898, x=[1.2000 0.8000].

Figure 3.8a shows the approaching of the two local optima from the starting points [1.5 0.7] and [2.6 1.2], using MSM method. Figure 3.8b shows the SMS method finding the two local optima from the same starting points. Compared with MSM, SMS need fewer iterations (number of triangles in the figure) to reach the optima, this is clearly presented in the result Figure 3.8. However, as the SMS algorithm required the evaluation of the centroid point in each it-




**(a)** MSM method.




**(b)** SMS method.

**Figure 3.8:** Demonstration of solving a 2-D problem using simplex methods. Two processes were given in both figures. One start from [1.5 0.7], the other start from [2.6 1.2].

eration, the total number of calling the object function is not necessarily lower. When starting from [1.2 0.7], MSM takes 13 iterations, 25 callings of the object function, while SMS only needed 7 iterations, and 21 callings to reach the optimum point. In this case, SMS takes fewer iteration and experiment calling to find the optima. When starting from [2.6 1.2], MSM takes 29 iterations, and 56 callings of the object function, while SMS takes 26 iterations, and 78 callings to reach the optima. In the [2.6 1.2] starting point case, even though SMS uses fewer iteration, the total number of calling the object function was higher.

**Investigating of the performance on higher dimension problems**

Nelder-Mead suffers from the 'curse of dimensionality'. Simplex methods do reasonably well in low numbers of dimensions (two or three dimensions). When it comes to higher dimension problems, the algorithm's performance is not as satisfactory. The SMS and MSM are further compared in higher dimension optimisation problem in this section.

In order to test the performance of both MSM and SMS on different dimension problems, the Rosenbrock function is used. It was first introduced by Rosenbrock [133], and it has been widely used as a test function for optimisation algorithms [124, 131, 134]. The function is chosen for the reason that it can be easily changed to multidimensional generalisations. The function is defined by:

$$f(x) = \sum_{i=1}^{n-1} [(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2] \tag{3.2.3}$$

The global minimum is at $f(1, 1 \cdots, 1) = 0$.

When n=2, Rosenbrock is a 2-D problem, the global minimum is inside a long, narrow, parabolic shaped flat valley.

Using default stopping criterion, [-1.5 1.5] as the search interval for reach dimension, origin as the starting point, 'fminsearchbnd' and 'SMSbnd' were tested on 2 to 10 dimensions Rosenbrock problem. The result is summarised in Table 3.7, with all the experiments stopped by 'MaxFunEvals' highlighted in orange colour. In the 'Optima found' column, Y=Yes and N=No.

**Table 3.7:** MSM and SMS performance on multi-dimension Rosenbrock functions, with all the experiments stopped by 'MaxFunEvals' (failed to reach optima) highlighted in orange colour.

| Problem dimension | MSM | | | SMS | | |
|---|---|---|---|---|---|---|
| | Iterations | funcCount | Optima found | Iterations | funcCount | Optima found |
| 2 | 79 | 150 | Y | 131 | 393 | Y |
| 3 | 101 | 190 | Y | 56 | 169 | N |
| 4 | 207 | 348 | Y | 266 | 800 | N |
| 5 | 347 | 553 | Y | 271 | 816 | N |
| 6 | 507 | 783 | Y | 206 | 622 | N |
| 7 | 745 | 1127 | Y | 171 | 518 | N |
| 8 | 1097 | 1600 | N | 233 | 705 | N |
| 9 | 1261 | 1800 | N | 152 | 463 | N |
| 10 | 1427 | 2000 | N | 239 | 725 | N |

As expected, MSM method performance drops as the dimension increase. From a 4-D to a 7-D problem, MSM is still able to find the optima, albeit with a huge

requirement on the calling of the object function. When the problem becomes 8-D or higher, the algorithm falls in reaching the global optimum within the funcCount maximum value $200 \times dimension$.

Comparing with MSM, SMS method's performance is not as satisfactory. It successfully found the optima in a 2-D problem. However, in 4-D problem, it did not reach an optimum within the maximum funcCount value 800. In all other dimension problems, the algorithm returns another local optimum worse than the optima given by MSM. This is because the SMS simplex follows the landscape more closely, which means the method is more easily distracted by other local optima.

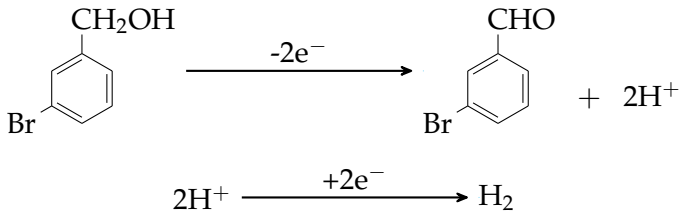
In conclusion, the simulator based optimisation indicated that MSM should be adopted for the self-optimising reactor. SMS method does have the advantage of using fewer iterations, however, it is the number of calling the object function matters to the self-optimising reactor as it is the most time-consuming step in our optimisation process.

### 3.2.2 Performance on self-optimising system

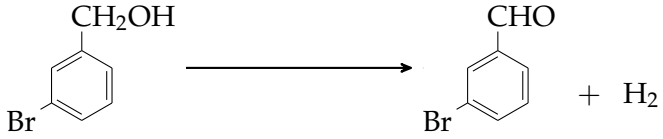
The electrochemical oxidation of 3-bromobenzyl alcohol (Scheme 3.2) was used to test the performance of the system with MSM optimisation approach. The reaction was selected since this reaction can provide a complex enough situation to test the optimisation algorithms.

**Anode:**

**Cathode:**




**Overall:**




**Scheme 3.2:** The electro-oxidation of 3-bromobenzyl alcohol.

A 3-D optimisation process is adopted in this section. The three factors and the response obtained from the GC measurements are given in Table 3.8, together with the lower and upper bounds of each factor. The general form of the landscape function is given in Equation 3.2.4.

$$Yield = f(I_{applied}, I_{applied}/I_{theoretical}, TotalFlowRate) \tag{3.2.4}$$

$I_{applied}$ is the electrical current applied to the Ammonite electrochemical reactor. $I_{theoretical}$ represents the electrical current required to achieve 100% yield according to Faraday's laws of electrolysis, assuming that no side reactions occur. The equation for $I_{theoretical}$ calculation is given in Equation 3.1.3 on page 45.

The test reactions were carried out in the Ammonite electrochemical reactor. The $CH_3CN$ solvent reactant streams (0.45 M 3-bromobenzyl alcohol + 0.05 M $Et_4NBF_4$ + 2 g/L 1-Hexanenitrile + 50 g/L $H_2O$) and (0.05 M $Et_4NBF_4$ + 4 g/L valeronitrile + 50 g/L $H_2O$) were pumped into the spiral micro-flow cell reactor, R, and then, to a gas-liquid separator, S. Two solutions were prepared with an identical concentration of $Et_4NBF_4$ (0.050 M) as the supporting electrolyte. The

**Table 3.8:** Factors and response of the landscape function for the reaction of the electro-oxidation of 3-bromobenzyl alcohol

| | Name | Low bound | Upper bound | Starting point |
|---|---|---|---|---|
| Factor 1 | $I_{applied}$ | 0.2 | 0.6 | 0.4 |
| Factor 2 | $I_{applied}/I_{theoretical}$ | 0.9 | 1.2 | 1.05 |
| Factor 3 | Total Flow Rate (ml/min) | 0.5 | 1.5 | 1 |
| Response | Yield | - | - | - |

concentration of the starting material (3-bromobenzyl alcohol) is 0.45 M in the first solution; and no starting material was added to the second solution. Based on the given factors for each experiment, the concentration of the starting material (3-bromobenzyl alcohol) was obtained using Equation 3.2.5. The variation of the concentration of starting material was achieved by changing the ratio of the two flow rates for the two solutions.

$$C_{startingmaterial} = \frac{I_{theoretical} \cdot 60 \cdot 1000}{nF \cdot FlowRate_{total}} = \frac{Factor1 \cdot 60 \cdot 1000}{Factor2 \cdot nF \cdot Factor3} \tag{3.2.5}$$

where $n$ is the number of electrons involved, $F$ is the Faraday constant, and the unit for $C_{startingmaterial}$ is M.

An internal standard method to evaluate the accuracy of the pumps is introduced in this experiment. 1-Hexanenitrile in solution 1, and valeronitrile in solution 2 were not directly involved in the reaction, and detectable by the GC methods used. The GC spectrum peaks of those two materials in the resultant solution should be in a linear relationship with the pump flow rate of that solu-

tion. This can provide an extra evaluation of the pump flow rate recorded by the automation software.

Based on the rig and software setting for Section 3.1.2, reconfiguring the rig the software was achieved in less than 4 hours. With boundary setting of the search domain and the starting point of each variable (factor) in Table 3.8, and the stopping criterion highlighted in Table 3.5, a 3-D optimisation of the oxidation of 3-bromobenzyl alcohol reaction was set to run on the automated self-optimisation flow chemistry rig. The process kept running for 24 hours and exit on the stopping criterion 1 *'the maximum coordinate-difference between the current best point, and the other points in the simplex is less than or equal to StepTolerance (TolX). **AND** the corresponding difference in function values is less than or equal to FunctionTolerance (TolFun).'*

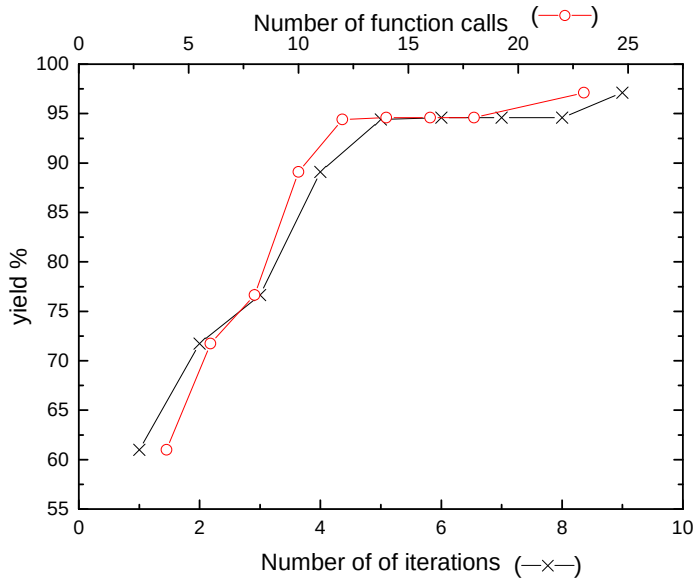



**Figure 3.9:** SMSIM optimisation of the yield. The yield is plotted as the number of iterations and the number of function calls as the optimisation process proceeds.

The algorithm decides to run 23 reactions, resulting that the best experimental

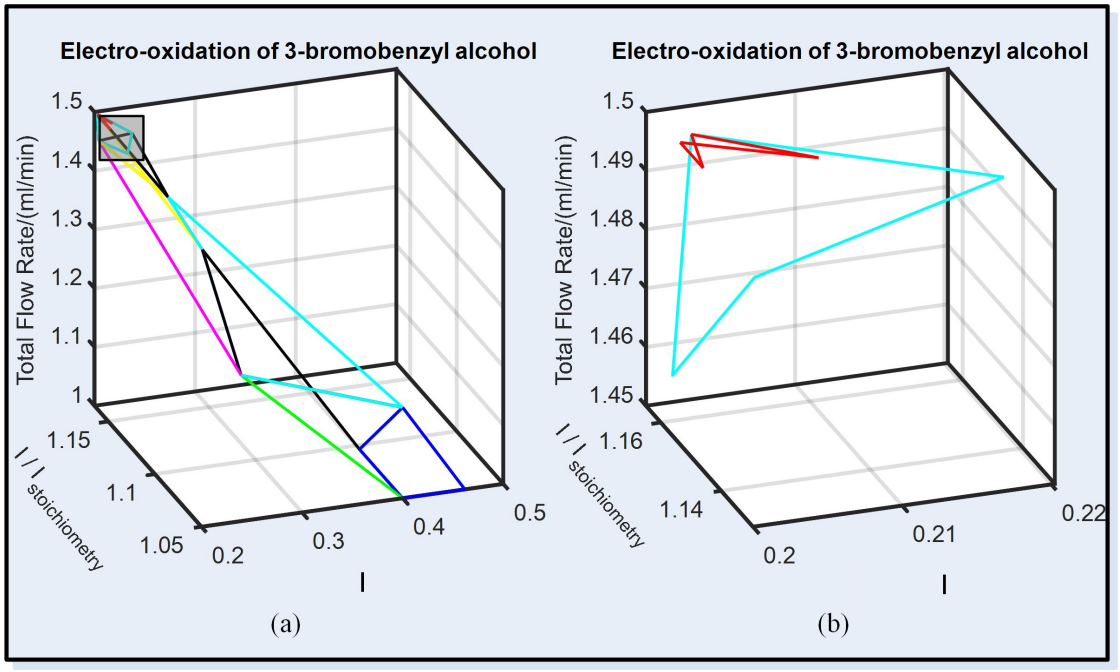



**Figure 3.10:** NMSIM optimisation of the yield of electro-oxidation of 3-bromobenzyl alcohol, with respect to the current, the electrical current ratio and the total flow rate of the inlet mixture (see Table 3.8). Factors 1, 2 and 3 are represented by x-, y- and z-axis. (a) The initial simplex is shown by a blue quadrangle. (b) Zoomed area of the top-left corner of (a). The last simplex is shown by a red quadrangle.

condition is at 0.202 A for the current, 1.163 for the current ratio and 1.495 ml/min for the total flow rate. The best yield obtained from the algorithm is 97.11%.

Figures 3.9 and 3.10 show the optimisation procedure using NMSIM with the D'Errico method for the treatment of boundary conditions. In Figures 3.10, the yield of the product increases steadily as the iterations of the algorithm increase. Starting from 60%, within four iterations (12 experiments), the yield of the product is already reaching 95%. This proves the efficiency of the Simplex algorithm. The simplex generated during this optimisation process is plotted

in Figure 3.9, with x-, y-, z-axis representing one of the three factors being optimised. As this is a 3-D problem, a four vertices simplex is used.

Successfully running a 3-D MSM optimisation indicated, not only that the MSM algorithm was well fitted to the automation software for flow chemistry rig, but also the great flexibility and reconfigurability of the software system developed for flow chemistry optimisation. Within hours of configuration, we were able to conduct a higher dimension optimisation on a new reaction. This shows the great discovery potential of the self-optimisation flow chemistry rig.

CHAPTER 4

# Conclusion and future work

In this work, we have developed an automated self-optimising continuous-flow reactor system to perform electrochemical reactions. Two model reactions, methoxylation of N-formylpyrrolidine and oxidation of 3-bromobenzyl alcohol were optimised based on different metrics for reaction success.

A MATLAB & LabVIEW based automation software was developed to conduct automation, control and monitoring of the flow chemistry rig. A server-client software structure was adopted to provide flexibility and reconfigurability to the software. The smooth transfer from a SNOBFIT based 2-D optimisation on methoxylation of N-formylpyrrolidine reaction to a Simplex method based 3-D optimisation on oxidation of 3-bromobenzyl alcohol indicated the system is a flexible, reliable control and automation software system.

SNOBFIT and Simplex optimisation methods were fitted to the automation software, and the whole system was tested both on multiple simulators and the rig itself. The pros and cons of each algorithm are clearly discussed supported by the simulation/experiment data.

The NMSIM algorithm comprises of a series of simple rules to modify the simplex repeatedly, and as such are ideally suited to an experimental procedure [64, 124, 127]. However, the capabilities of the NMSIM algorithm to adapt the control equipments' (pump, power supply, *ect*.) accuracy is restricted, and the lack of global search functions can result in poor confidence in the located optimum. By comparison, SNOBFIT combines both global and local search elements to provide a powerful tool for the global optimum search. This discrete value optimisation algorithm can define the minimum search step-size on each dimension, which adapts well to experimental equipment accuracy. However, both simulator and experiments indicated it requires more iterations, thus higher number of experiments to reach the optimum.

This work has shown how the efficiency of reactions and chemical processes can be improved by a highly automatic, self-optimising continuous-flow reactor. It demonstrated use of continuous-flow as a tool for discovery in chemistry, for the purpose of making chemistry greener.

## 4.1 Further work on rig

We now have a fully functioning software application to optimise electrochemical reactions using the Ammonite as the reactor and GC as the on-line analysis tool. We shall have opportunities to optimise a variety of electrochemical reactions in the future. On the other hand, it is required further studies on how to run these optimisation algorithms efficiently, namely to find the optimal conditions with fewer numbers of experiments. Each optimisation algorithm re-

quires some inputs to define how to run optimisation processes. The input parameters can be the range of each factor, the grid size of each factor, the choices of the starting point for the Simplex and Super-Modified Simplex Algorithms. More importantly, the uncertainty of the measured response resulting from the analytical instrument has significant effects on the performance of the optimisation algorithms. It would be useful to develop a practical guidance on how to choose the parameters for optimisation algorithms and on how to relate the experimental uncertainties to these parameters.

## 4.2 Alternative target functions

The use of automated continuous-flow self-optimising reactor system represents a significant step forward in the used of optimisation algorithms for chemical reactions. However, it is not merely the maximum yield of a product, which can be optimised. The target function could also be optimised for different criteria, such as maximising the ration of two products, minimising the production of an unwanted by-product, or even minimising the E-factor of the reaction, which is particularly interesting in the context of Green Chemistry [31, 57]. We have demonstrated the using of different target functions in the two model reactions. Further exploring of this aspect is needed.

As well as the ability to optimise a reaction, the self-optimisation approach to a chemical reaction has the advantage of discovering unexpected and original synthetic reactions. A self-optimising reactor was used by Amara et al. [103] as a tool for targeting known and unknown materials in the continuous reac-

tions of aniline, dimethyl carbonate (DMC) and tetrahydrofuran (THF) in supercritical $CO_2$ on $\gamma$-$Al_2O_3$, which has already been identified that could form multiple products in parallel. The results demonstrated that self-optimising reactors enable chemists to switch the selectivity to different products in a chemically complex system.

## 4.3 Alternative analysis technique

Although (low level of manual control and monitoring needed), the time required for self-optimisation can be quite lengthy because of the time required for GLC analysis. Skilton et al. reported the use of automated continuous flow reactors with real-time online Fourier transform infrared spectroscopy analysis to enable rapid optimisation of reaction yield using a self-optimising feedback algorithm [104]. The results indicated that the high speed of the optimisation itself provides major savings in material and permits a wider exploration of parameter space than would be possible with GLC alone [104].

The methods of online analysis can be varied in the future work of this thesis. Investigations of applying real-time monitoring analysis technique, FTIR analysis, for example, can result in a more rapid optimisation process, which will lead to significant material savings. The real-time feedback analysis can also permit a more detailed study of the response space/surface of an unknown reaction, which will lead to a more in-depth understanding of the reaction.

## 4.4 Alternative solvents

Solvent can be a very influential fact of a reaction [36]. Ever since the birth of Green Chemistry, solvents have occupied a central place in the effort to make chemical processes greener [57]. Pharmaceutical companies are increasingly aiming to reduce their environmental impact due to solvent choice [135]. Supercritical solvents are highly compressed and/or heated gases that are beyond the critical point. It is a greener alternative to conventional solvents [102]. Using supercritical solvents for the automated self-optimisation continuous-flow reactor system can be beneficial.

Streng et al. used an automated self-optimising reactor to determine the optimal reaction conditions for the synthesis of N-alkylated heterocycles. Targeted N-methylpiperidine, using the self-optimisation approach with SNOBFIT as the optimising algorithm and GC analysis as the analytical tool, the temperature and the flow rate of the reaction were optimised in both the presence and absence of Supercritical Carbon Dioxide ($scCO_2$). The result shows higher yields were obtained in the presence of $CO_2$ than in its absence [102].

# Bibliography


[1] P. T. Anastas and J. C. Warner. *Green Chemistry: Theory and Practice*. New York: Oxford University Press, 1998.

[2] Sebastian Steiner et al. "Organic synthesis in a modular robotic system driven by a chemical programming language". In: *Science* (2018), eaav2211.

[3] Andrea Adamo et al. "On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system". In: *Science* 352.6281 (2016), pp. 61–67.

[4] Norbert Kockmann et al. "Enabling continuous-flow chemistry in microstructured devices for pharmaceutical and fine-chemical production". In: *Chem.–Eur. J.* 14.25 (2008), pp. 7470–7477.

[5] Charlotte Wiles and Paul Watts. "Continuous flow reactors, a tool for the modern synthetic chemist". In: *Eur. J. Org. Chem.* 2008.10 (2008), pp. 1655–1671.

[6] Brian K. H. Yen et al. "A microfabricated gas–liquid segmented flow reactor for high-temperature synthesis: the case of CdSe quantum dots". In: *Angew. Chem. Int. Ed.* 44.34 (2005), pp. 5447–5451.

[7] Saif A. Khan et al. "Microfluidic synthesis of colloidal silica". In: *Langmuir* 20.20 (2004), pp. 8604–8611.

[8] Daniel M. Ratner et al. "Microreactor-based reaction optimization in organic chemistry—glycosylation as a challenge". In: *Chem. Commun.* 5 (2005), pp. 578–580.

[9] Andrew J. Demello. "Control and detection of chemical reactions in microfluidic systems." In: *Nature* 442.7101 (2006), pp. 394–402.

[10] Julio C. Pastre, Duncan L. Browne and Steven V. Ley. "Flow chemistry syntheses of natural products". In: *Chem. Soc. Rev.* 42.23 (2013), pp. 8849–8869.

[11] Brian P. Mason et al. "Greener approaches to organic synthesis using microreactor technology". In: *Chem. Rev.* 107.6 (2007), pp. 2300–2318.

[12] Ben Walsh et al. "The automation of continuous reactions in supercritical CO 2: the acid-catalysed etherification of short chain alcohols". In: *Green Chem.* 7.6 (2005), pp. 456–463.

[13] Richard A. Bourne et al. "Maximising opportunities in supercritical chemistry: the continuous conversion of levulinic acid to $\gamma$-valerolactone in CO2". In: *Chem. Commun.* 44 (2007), pp. 4632–4634.

[14] James G. Stevens et al. "Real-time product switching using a twin catalyst system for the hydrogenation of furfural in supercritical CO2". In: *Angew. Chem.* 122.47 (2010), pp. 9040–9043.

[15] James G. Stevens, Richard A. Bourne and M. Poliakoff. "The continuous self aldol condensation of propionaldehyde in supercritical carbon dioxide: a highly selective catalytic route to 2-methylpentenal". In: *Green Chem.* 11.3 (2009), pp. 409–416.

[16] Peter Licence et al. "Selective monoprotection of 1, n-terminal diols in supercritical carbon dioxide: a striking example of solvent tunable desymmetrization". In: *J. Am. Chem. Soc.* 127.1 (2005), pp. 293–298.

[17] Andrew J. Parrott et al. "The continuous acid-catalysed etherification of aliphatic alcohols using stoichiometric quantities of dialkyl carbonates". In: *Org. Process Res. Dev.* 14.6 (2010), pp. 1420–1426.

[18] Peter N. Gooden et al. "Continuous acid-catalyzed methylations in supercritical carbon dioxide: comparison of methanol, dimethyl ether and dimethyl carbonate as methylating agents". In: *Org. Process Res. Dev.* 14.2 (2010), pp. 411–416.

[19] H. Perry and D. W. Green. *Perry's Chmeical Engineers' Handbook*. McGraw-Hill, 1999, pp. 2085–2095.

[20] W. B. Krantz. *Scaling analysis in modeling transport and reaction processes: a systematic approach to model building and the art of approximation*. John Wiley & Sons, 2007.

[21] Oliver Flögel et al. "Microreactor Synthesis of $\beta$-Peptides". In: *Angew. Chem. Int. Ed.* 45.42 (2006), pp. 7000–7003.

[22] Tatsuya Kawaguchi et al. "Room-temperature Swern oxidations by using a microscale flow system". In: *Angew. Chem. Int. Ed.* 44.16 (2005), pp. 2413–2416.

[23] Ashok K. Sharma and Daniel Swern. "Trifluoroacetic anhydride: A new activating agent for dimethyl sulfoxide in the synthesis of iminosulfuranes". In: *Tetrahedron Lett.* 15.16 (1974), pp. 1503–1506.

[24] Tahseen Razzaq and C. O. Kappe. "Continuous flow organic synthesis under high-temperature/pressure conditions". In: *Chem. Asian J.* 5.6 (2010), pp. 1274–1289.

[25] Paul Anastas and Nicolas Eghbali. “Green chemistry: principles and practice”. In: *Chem. Soc. Rev.* 39.1 (2010), pp. 301–312.

[26] Stephen G. Newman and Klavs F. Jensen. “The role of flow in green chemistry and engineering”. In: *Green Chem.* 15.6 (2013), pp. 1456–1472.

[27] N. G. Anderson. “Using continuous processes to increase production”. In: *Org. Process Res. Dev.* 16.5 (2012), pp. 852–869.

[28] Charlotte Wiles and Paul Watts. “Recent advances in micro reaction technology”. In: *Chem. Commun.* 47.23 (2011), pp. 6512–6535.

[29] Jun-ichi Yoshida. *Flash chemistry: Fast organic synthesis in microsystems*. John Wiley & Sons, 2008.

[30] Charlotte Wiles and Paul Watts. *Micro reaction technology in organic synthesis*. CRC Press, 2016.

[31] Roger A. Sheldon. “The E factor: fifteen years on”. In: *Green Chem.* 9.12 (2007), pp. 1273–1283.

[32] R. A. Sheldon. “E factors, green chemistry and catalysis: an odyssey”. In: *Chem. Commun.* 29 (2008), pp. 3352–3365.

[33] Wen-Chung Shieh, Steven Dell and Oljan Repic. “1, 8-Diazabicyclo [5.4. 0] undec-7-ene (DBU) and microwave-accelerated green chemistry in methylation of phenols, indoles, and benzimidazoles with dimethyl carbonate”. In: *Org. Lett.* 3.26 (2001), pp. 4279–4281.

[34] Ulf Tilstam. “A Continuous Methylation of Phenols and N, H-Heteroaromatic Compounds with Dimethyl Carbonate”. In: *Org. Process Res. Dev.* 16.12 (2012), pp. 1974–1978.

[35] Dana Kralisch et al. “Transfer of the epoxidation of soybean oil from batch to flow chemistry guided by cost and environmental issues”. In: *ChemSusChem* 5.2 (2012), pp. 300–311.

[36] Ryan Alan Skilton. “Chemical reactions in continuous flow”. PhD thesis. University of Nottingham, 2014.

[37] Food and Drug Administration. *Strategic Plan for Preventing and Mitigating Drug Shortages*. 2013. URL: `http://www.fda.gov/downloads/Drugs/DrugSafety/DrugShortages/UCM372566.pdf` (visited on 12/12/2018).

[38] Spencer D. Schaber et al. “Economic analysis of integrated continuous and batch pharmaceutical manufacturing: a case study”. In: *Ind. Eng. Chem. Res.* 50.17 (2011), pp. 10083–10092.

[39] Concepción Jiménez-González et al. “Key green engineering research areas for sustainable manufacturing: a perspective from pharmaceutical and fine chemicals manufacturers”. In: *Org. Process Res. Dev.* 15.4 (2011), pp. 900–911.

[40] David J. C. Constable et al. “Key green chemistry research areas—a perspective from pharmaceutical manufacturers”. In: *Green Chem.* 9.5 (2007), pp. 411–420.

[41] B. S. Kwak. In: *Chem. Today* (2003 January-February).

[42] Dominique M. Roberge et al. “Microreactor technology and continuous processes in the fine chemical and pharmaceutical industry: is the revolution underway?” In: *Org. Process Res. Dev.* 12.5 (2008), pp. 905–910.

[43] Munawwer Rasheed and Thomas Wirth. “Intelligent Microflow: Development of Self-Optimizing Reaction Systems”. In: *Angew. Chem. Int. Ed.* 50.2 (2011), pp. 357–358.

[44] Sameer K. Ajmera et al. “Microfabricated differential reactor for heterogeneous gas phase catalyst testing”. In: *J. Catal.* 209.2 (2002), pp. 401–412.

[45] Tamara M. Floyd, Martin A. Schmidt and K. F. Jensen. “Silicon micromixers with infrared detection for studies of liquid-phase reactions”. In: *Ind. Eng. Chem. Res.* 44.8 (2005), pp. 2351–2358.

[46] Jonathan P. McMullen et al. “An Integrated Microreactor System for Self-Optimization of a Heck Reaction: From Micro-to Mesoscale Flow Systems”. In: *Angew. Chem. Int. Ed.* 49.39 (2010), pp. 7076–7080.

[47] Fernando Benito-Lopez et al. “Optical fiber-based on-line UV/Vis spectroscopic monitoring of chemical reaction kinetics under high pressure in a capillary microreactor”. In: *Chem. Commun.* 22 (2005), pp. 2857–2859.

[48] Heiko Lange et al. “A breakthrough method for the accurate addition of reagents in multi-step segmented flow processing”. In: *Chem. Sci.* 2.4 (2011), pp. 765–769.

[49] Sergey Mozharov et al. “Improved method for kinetic studies in microreactors using flow manipulation and noninvasive Raman spectrometry”. In: *J. Am. Chem. Soc.* 133.10 (2011), pp. 3601–3608.

[50] Helen Song and R. F. Ismagilov. “Millisecond kinetics on a microfluidic chip using nanoliters of reagents”. In: *J. Am. Chem. Soc.* 125.47 (2003), pp. 14613–14619.

[51] Luisa Ciobanu et al. “Measuring reaction kinetics by using multiple microcoil NMR spectroscopy”. In: *Angew. Chem. Int. Ed.* 42.38 (2003), pp. 4669–4672.

[52] Christopher J. Welch et al. “Online analysis of flowing streams using microflow HPLC”. In: *Org. Process Res. Dev.* 13.5 (2009), pp. 1022–1025.

[53] Duncan L. Browne et al. "Continuous flow reaction monitoring using an on-line miniature mass spectrometer". In: *Rapid Commun. Mass Spectrom.* 26.17 (2012), pp. 1999–2010.

[54] Patrick L. Mills and J. F. Nicole. "Multiple automated reactor systems (MARS). 1. A novel reactor system for detailed testing of gas-phase heterogeneous oxidation catalysts". In: *Ind. Eng. Chem. Res.* 44.16 (2005), pp. 6435–6452.

[55] Tracy J. Benson et al. "Elucidation of the catalytic cracking pathway for unsaturated mono-, di-, and triacylglycerides on solid acid catalysts". In: *J. Mol. Catal. A: Chem.* 303.1-2 (2009), pp. 117–123.

[56] Rachel Herzig-Marx et al. "Infrared spectroscopy for chemically specific sensing in silicon-based microreactors". In: *Anal. Chem.* 76.21 (2004), pp. 6476–6483.

[57] Xue Han and Martyn Poliakoff. "Continuous reactions in supercritical carbon dioxide: problems, solutions and possible ways forward". In: *Chem. Soc. Rev.* 41.4 (2012), pp. 1428–1436.

[58] John R. Goodell et al. "Development of an automated microfluidic reaction platform for multidimensional screening: reaction discovery employing bicyclo [3.2. 1] octanoid scaffolds". In: *J. Org. Chem.* 74.16 (2009), pp. 6169–6180.

[59] Eduardo Garcia-Egido et al. "Synthesis and analysis of combinatorial libraries performed in an automated micro reactor system". In: *Lab. Chip* 3.2 (2003), pp. 73–76.

[60] S. Krishnadasan, R. J. C. Brown et al. "Intelligent routes to the controlled synthesis of nanoparticles". In: *Lab. Chip* 7.11 (2007), pp. 1434–1441.

[61] Atsushi Sugimoto et al. "An automated-flow microreactor system for quick optimization and production: application of 10-and 100-gram order productions

of a matrix metalloproteinase inhibitor using a Sonogashira coupling reaction". In: *Tetrahedron Lett.* 50.46 (2009), pp. 6364–6367.

[62] Charlotte Wiles and Paul Watts. "Continuous flow reactors: a perspective". In: *Green Chem.* 14.1 (2012), pp. 38–54.

[63] Andrew J. Parrott et al. "Self-optimizing continuous reactions in supercritical carbon dioxide". In: *Angew. Chem. Int. Ed.* 50.16 (2011), pp. 3788–3792.

[64] Richard A. Bourne et al. "Adaptive process optimization for continuous methylation of alcohols in supercritical carbon dioxide". In: *Org. Process Res. Dev.* 15.4 (2011), pp. 932–938.

[65] Jonathan P. McMullen and Klavs F. Jensen. "An automated microfluidic system for online optimization in chemical synthesis". In: *Org. Process Res. Dev.* 14.5 (2010), pp. 1169–1176.

[66] Jason S. Moore and K. F. Jensen. "Automated multitrajectory method for reaction optimization in a microfluidic system using online IR analysis". In: *Org. Process Res. Dev.* 16.8 (2012), pp. 1409–1415.

[67] K. Poscharny et al. "Machine assisted reaction optimization: A self-optimizing reactor system for continuous-flow photochemical reactions". In: *Tetrahedron* 74.25 (2018), pp. 3171–3175.

[68] Zhenpeng Zhou, Xiaocheng Li and R. N. Zare. "Optimizing chemical reactions with deep reinforcement learning". In: *ACS Cent. Sci.* 3.12 (2017), pp. 1337–1344.

[69] Artur M. Schweidtmann et al. "Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives". In: *Chem. Eng. J.* 352 (2018), pp. 277–282.

[70] Anne-Catherine Bédard et al. "Reconfigurable system for automated optimization of diverse chemical reactions". In: *Science* 361.6408 (2018), pp. 1220–1225.

[71] Derek Pletcher, Robert A. Green and R. C. D. Brown. "Flow electrolysis cells for the synthetic organic chemistry laboratory". In: *Chem. Rev.* 118.9 (2017), pp. 4573–4591.

[72] Bernardo A. Frontana-Uribe et al. "Organic electrosynthesis: a promising green methodology in organic chemistry". In: *Green Chem.* 12.12 (2010), pp. 2099–2119.

[73] Michael Faraday. *Experimental researches in electricity*. Vol. 1. Read Books Ltd, 2016.

[74] Tatsuya Shōno. *Electroorganic synthesis*. Academic Pr, 1991.

[75] Ole Hammerich and Bernd Speiser. *Organic electrochemistry: revised and expanded*. CRC Press, 2015.

[76] Jun-ichi Yoshida et al. "Modern strategies in electroorganic synthesis". In: *Chem. Rev.* 108.7 (2008), pp. 2265–2299.

[77] Hermann Kolbe. "Untersuchungen über die Elektrolyse organischer Verbindungen". In: *Justus Liebigs Ann. Chem.* 69.3 (1849), pp. 257–294.

[78] K. F. Jensen. "Microreaction engineering—is small better?" In: *Chem. Eng. Sci.* 56.2 (2001), pp. 293–303.

[79] Thomas Wirth. *Microreactors in organic chemistry and catalysis*. John Wiley & Sons, 2013.

[80] Daisuke Horii et al. "Self-supported paired electrosynthesis of 2, 5-dimethoxy-2, 5-dihydrofuran using a thin layer flow cell without intentionally added supporting electrolyte". In: *Electrochem Commun.* 7.1 (2005), pp. 35–39.

[81] Daisuke Horii et al. "Self-supported methoxylation and acetoxylation electrosynthesis using a simple thin-layer flow cell". In: *J. Electrochem. Soc.* 153.8 (2006), pp. D143–D147.

[82] Daisuke Horii, Toshio Fuchigami and Mahito Atobe. "A new approach to anodic substitution reaction using parallel laminar flow in a micro-flow reactor". In: *J. Am. Chem. Soc.* 129.38 (2007), pp. 11692–11693.

[83] Daisuke Horii et al. "A novel electrosynthetic system for anodic substitution reactions by using parallel laminar flow in a microflow reactor". In: *Chem. Eur. J.* 14.33 (2008), pp. 10382–10387.

[84] Stuart M. MacDonald et al. "Electrochemical processes at a flowing organic solvent aqueous electrolyte phase boundary". In: *Electrochem. Commun.* 9.8 (2007), pp. 2105–2110.

[85] Jun-ichi Yoshida and Seiji Suga. "Basic concepts of "cation pool" and "cation flow" methods and their applications in conventional and combinatorial organic synthesis". In: *Chem.–Eur. J.* 8.12 (2002), pp. 2650–2658.

[86] Jun-ichi Yoshida et al. "Recent developments in the "cation pool" method". In: *J. Synth. Org. Chem Jpn.* 71.11 (2013), pp. 1136–1144.

[87] Jun-ichi Yoshida et al. "Reaction Integration Using Electrogenerated Cationic Intermediates". In: *Bull. Chem. Soc. Jpn.* 88.6 (2015), pp. 763–775.

[88] Robert A. Green et al. "A microflow electrolysis cell for laboratory synthesis on the multigram scale". In: *Org. Process Res. Dev.* 19.10 (2015), pp. 1424–1427.

[89] Robert A. Green et al. "An extended channel length microflow electrolysis cell for convenient laboratory synthesis". In: *Electrochem. Commun.* 73 (2016), pp. 63–66.

[90] Ping He et al. "Self-Supported and Clean One-Step Cathodic Coupling of Activated Olefins with Benzyl Bromide Derivatives in a Micro Flow Reactor". In: *Angew. Chem.* 118.25 (2006), pp. 4252–4255.

[91] Roberto Horcajada et al. "Microflow electroorganic synthesis without supporting electrolyte". In: *Chem. Commun.* 10 (2005), pp. 1303–1305.

[92] Seiji Suga et al. "Electrochemical combinatorial organic syntheses using microflow systems". In: *Qsar Comb. Sci.* 24.6 (2005), pp. 728–741.

[93] Siegfried R. Waldvogel and B. Janza. "Renaissance of electrosynthetic methods for the construction of complex molecules". In: *Angew. Chem. Int. Ed.* 53.28 (2014), pp. 7122–7123.

[94] Derek Pletcher, Zhong-Qun Tian and David Williams. *Developments in electrochemistry: science inspired by Martin Fleischmann*. John Wiley & Sons, 2014.

[95] Evan J. Horn, Brandon R. Rosen and P. S. Baran. "Synthetic organic electrochemistry: an enabling and innately sustainable method". In: *ACS Cent. Sci.* 2.5 (2016), pp. 302–308.

[96] Marcelo Navarro. "Recent advances in experimental procedures for electroorganic synthesis". In: *Curr. Opin. Chem. Biol.* 2.1 (2017), pp. 43–52.

[97] Ryan A. Skilton et al. "Remote-controlled experiments with cloud chemistry". In: *Nat. Chem.* 7 (2015), pp. 1–5.

[98] Richard A. Bourne. "Automated reactions in supercritical carbon dioxide". PhD thesis. University of Nottingham, 2008.

[99] Ben Walsh. "Automated Reactions in Supercritical Carbon Dioxide". PhD thesis. Univerisity of Nottingham, 2004.

[100] Andrew John Parrott. "Continuous clean etherification". PhD thesis. University of Nottingham, 2011.

[101] Emilia S. Streng. "Carbon Dioxide as a Reagent and Solvent". PhD thesis. University of Nottingham, 2017.

[102] Emilia S. Streng et al. "Continuous N-alkylation reactions of amino alcohols using $\gamma$-Al2O3 and supercritical CO2: unexpected formation of cyclic ureas and urethanes by reaction with CO2". In: *Beilstein J. Org. Chem.* 13 (2017), p. 329.

[103] Zacharias Amara et al. "Automated Serendipity with Self-Optimizing Continuous-Flow Reactors". In: *Eur. J. Org. Chem.* 2015.28 (2015), pp. 6141–6145.

[104] Ryan A. Skilton et al. "Real-time feedback control using online attenuated total reflection Fourier transform infrared (ATR FT-IR) spectroscopy for continuous flow optimization and process knowledge". In: *Appl. Spectrosc.* 67.10 (2013), pp. 1127–1131.

[105] Denis N. Jumbam et al. "The effect of self-optimisation targets on the methylation of alcohols using dimethyl carbonate in supercritical CO2". In: *J. Flow. Chem.* 2.1 (2012), pp. 24–27.

[106] Richard J. Ingham et al. "A systems approach towards an intelligent and self-controlling platform for integrated continuous reaction sequences". In: *Angew. Chem. Int. Ed.* 54.1 (2015), pp. 144–148.

[107] Nikolay Cherkasov et al. "OpenFlowChem–a platform for quick, robust and flexible automation and self-optimisation of flow chemistry". In: *React. Chem. Eng.* 3.5 (2018), pp. 769–780.

[108] Victor Sans and Leroy Cronin. "Towards dial-a-molecule by integrating continuous flow, analytics and self-optimisation". In: *Chem. Soc. Rev.* 45.8 (2016), pp. 2032–2043.

[109] Brandon J. Reizman and K. F. Jensen. "Feedback in flow for accelerated reaction development". In: *Acc. Chem. Res.* 49.9 (2016), pp. 1786–1796.

[110] Victor Sans et al. "A self optimizing synthetic organic reactor system using real-time in-line NMR spectroscopy". In: *Chem Sci* 6.2 (2015), pp. 1258–1264.

[111] Brandon J. Reizman and K. F. Jensen. "An automated continuous-flow platform for the estimation of multistep reaction kinetics". In: *Org. Process Res. Dev.* 16.11 (2012), pp. 1770–1782.

[112] P. N. Bartlett. *Bioelectrochemistry: fundamentals, experimental techniques and applications*. John Wiley & Sons, 2008.

[113] Jason R. Hyde et al. "Continuous hydrogenation reactions in supercritical CO 2". In: *Green Chem.* 7.5 (2005), pp. 357–361.

[114] Waltraud Huyer and Arnold Neumaier. "SNOBFIT–stable noisy optimization by branch and fit". In: *ACM TOMS* 35.2 (2008), p. 9.

[115] Arnold Neumaier. `http://www.mat.univie.ac.at/~neum/software/snobfit/`. Accessed: 2018-07-31.

[116] Arnold Neumaier. `http://www.mat.univie.ac.at/~neum/software/minq/`. Accessed: 2018-07-31.

[117] Q. Y. Duan, Vijai K. Gupta and S. Sorooshian. "Shuffled complex evolution approach for effective and efficient global minimization". In: *J. Optimiz. Theory App.* 76.3 (1993), pp. 501–521.

[118] Jekaterina Kuleshova et al. "A simple and inexpensive microfluidic electrolysis cell". In: *Electrochim. Acta* 56.11 (2011), pp. 4322–4326.

[119] Robert A. Green, Richard C. D. Brown and Derek Pletcher. "Understanding the performance of a microfluidic electrolysis cell for routine organic electrosynthesis". In: *J. Flow. Chem.* 5.1 (2015), pp. 31–36.

[120] Jekaterina Kuleshova et al. "The methoxylation of N-formylpyrrolidine in a microfluidic electrolysis cell for routine synthesis". In: *Electrochim. Acta* 69 (2012), pp. 197–202.

[121] Henning Lund and M. M. Baizer. *Organic electrochemistry: an introduction and a guide*. Dekker New York, 1991.

[122] James Utley. "Trends in organic electrosynthesis". In: *Chem. Soc. Rev.* 26.3 (1997), pp. 157–167.

[123] W. G. R. F. R. Spendley, George R. Hext and F. R. Himsworth. "Sequential application of simplex designs in optimisation and evolutionary operation". In: *Technometrics* 4.4 (1962), pp. 441–461.

[124] John A. Nelder and Roger Mead. "A simplex method for function minimization". In: *Comput. J.* 7.4 (1965), pp. 308–313.

[125] Inc The MathWorks, ed. *Optimizing Nonlinear Functions*. `https://www.mathworks.com/help/matlab/math/optimizing-nonlinear-functions.html`. Accessed: 2018-07-17.

[126] John D'Errico. *Bound constrained optimization using fminsearch*. Ed. by MathWorks. `https://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd-fminsearchcon`. Accessed: 2018-07-17. 6th Feb. 2012.

[127] M. Wm Routh, P. A. Swartz and M. B. Denton. "Performance of the super modified simplex". In: *Anal. Chem.* 49.9 (1977), pp. 1422–1428.

[128] P. F. A. Van Der Wiel. "Improvement of the super-modified simplex optimization procedure". In: *Anal. Chim. Acta* 122.4 (1980), pp. 395–402.

[129] Edward Morgan, Kenneth W. Burton and G. Nickless. "Optimization using the super-modified simplex method". In: *Chemom. Intell. Lab Syst.* 8.2 (1990), pp. 97–107.

[130] L. R. Parker Jr, M. R. Cave and R. M. Barnes. "Comparison of simplex algorithms". In: *Anal. Chim. Acta* 175 (1985), pp. 231–237.

[131] Margarida F. Cardoso, Romualdo L. Salcedo and S. F. De Azevedo. "The simplex-simulated annealing approach to continuous non-linear optimization". In: *Comput. Chem. Eng.* 20.9 (1996), pp. 1065–1080.

[132] MathWorks, ed. *Tolerances and Stopping Criteria*. `https://www.mathworks.com/help/optim/ug/tolerances-and-stopping-criteria.html`. Accessed: 2018-07-26.

[133] HoHo Rosenbrock. "An automatic method for finding the greatest or least value of a function". In: *Comput. J.* 3.3 (1960), pp. 175–184.

[134] Roger Fletcher and M. J. Powell. "A rapidly convergent descent method for minimization". In: *Comput. J.* 6.2 (1963), pp. 163–168.

[135] Denis Prat et al. "CHEM21 selection guide of classical-and less classical-solvents". In: *Green Chem.* 18.1 (2015), pp. 288–296.

# Part II

# Applying FTIR Imaging to Address Challenges in Plastic Recycling

CHAPTER 5

# Introduction

Plastics are inexpensive, durable and one of the most widely-used materials, with diverse applications including packaging, automotive and general merchandising [1]. A major problem in the use of such polymer materials is that the majority of plastics have been developed specifically to persist degradation or natural decay process and they can take up to millions of years to degradate naturally [2]. The accumulation of plastics has begun to negatively impact the natural environment and create problems for plants, wildlife and even human populations. Plastic pollution is one of the major pollution problems in the environment and has increasing public and worldwide attention. Plastic recycling is an important strategy in order to combat the problem of plastic pollution. The re-use/recycling of plastic is not only environmentally friendly but also economically beneficial. In the European Union, the end-of-life vehicles (ELV) and waste electrical and electronic equipment (WEEE) directives have high requirements for the amount of plastic requiring recycling [3]. The first plastic recycling mill for Waste Techniques was built in Conshohocken, Pennsylvania,

and began working in 1972 [4]. It took several years and a concerted effort for the public to embrace the recycling habit. Plastic recycling is unlike glass or metal recycling processes due to the greater number of steps involved and the use of dyes, fillers and other additives used in 'virgin' plastics. The plastic recycling process begins with sorting the various items by their resin content. Once sorted, the plastics are chopped into small pieces, cleaned, melted down and compressed into pellets. Once pelletised, the recycled plastic pellets are ready to reuse and fashion into new and completely different products. It is important to note that recycled plastic is rarely used to create the same or identical plastic item of its former self, because the properties of recycled plastics are almost never as good as the virgin plastic.

**Figure 5.1:** Staff at Liansamlet Household Waste and Recycling Centre sorting through plastic. Reproduced from [5].

One of the many challenges for plastic recycling, is the identification of the wide range of plastic materials used. Currently, the state of art identification and quality classification of polymers is based on human hand sorting oper-

ation, as shown in Figure 5.1. This is time-consuming, and only practical to most domestic wastes, which typically contain five types of polymers (polyethylene terephthalate, polyethylene, polyvinyl chloride, polypropylene and polystyrene) [6] in large quantities. Advanced technical products such as automobile part productions, electrical and electronic equipment use a much wider range of plastic and filler/additives to gain different physical/chemical features, which is difficult to identify using current sorting techniques. As a result, the recycled plastic products have a high level of quality variation. The high variation of performance has been a major barriers to the successful market penetration of recycled plastics [3] for advanced technical products, and much plastic recycling is targeted at low-level applications.

Fourier transform infrared (FTIR) chemical imaging is a strongly emerging technology that is being increasingly applied to material investigations. The use of IR spectroscopy in polymer industries has been widely implemented [7–12]. However, majority of research has focussed on manufacturing, as well as the physical/chemical properties. This Chapter aims to explore the potential of FTIR spectroscopy to aid and improve plastic recycling. A quantitative approach is made to evaluate the detailed composition of plastics during reprocessing. A feasibility study using FTIR imaging for the rapid and label-free classification/evaluation of recycled plastic were conducted. The results indicated several promising ways to apply FTIR spectroscopy on automotive plastic recycling for high-end applications.

## 5.1 Application of IR spectroscopy to study polymers

Infrared spectroscopy has a long tradition and remains one of the most widely used spectroscopic techniques in the analysis and characterisation of polymers [9].

**Generic identification** is one of the earliest applications of IR spectroscopy on polymer studies. In the generic identification method, the polymer or copolymer is considered as a collection of functional groups or sub-structural units, for which the observed absorption bands are assigned using well-established group frequency correlation tables [7, 8]. A good example of generic identification is provide by Krimm et al. [12]. IR spectra of polyethylene (PE) were first time collected over 70-3000 $cm^{-1}$ wavenumber range. The assignments of the fundamentals are made with the help of a group theory analysis. The assignment of the $CH_2$ wagging mode to a weak band at 1369 $cm^{-1}$ was discussed in detail and especially in terms of new evidence from the spectra of n-paraffins, both as single crystals and as polycrystalline aggregates. The frequencies of the observed band maxima, their approximate relative intensities, and their dichroism with respect to the direction of stretch are given in Figure 5.2. This table was used to conduct a knowledge based spectra analysis in Section 10.1 on page 219.

| Frequency, $cm^{-1}$ | Intensity | Polarization | Assignment[a] |
|---|---|---|---|
| 2959 | w(sh) | ··· | $\nu_a(CH_3)$ |
| 2925 | vs | $\sigma$ | $\nu_a(CH_2)$[c] |
| 2874[b] | vw | ··· | $\nu_s(CH_3)$ |
| 2853 | s | $\sigma$ | $\nu_s(CH_2)$[d] |
| 2640 | w | $\sigma$ | |
| 2295 | vw | $\sigma$ | |
| 2130 | vvw | $\sigma$ | |
| 2010 | vw | $\pi$ | 1303+721=2024 |
| 1890 | vw | $\sigma$ | |
| 1805 | vvw | $\pi$ | |
| 1710 | vw | $\sigma$ | |
| 1470[b] | s | $\sigma$ | $\delta(CH_2)$ |
| 1460 | s | $\sigma$ | |
| 1456[b] | vw | ··· | $\delta_a(CH_3)$ |
| 1375 | m | $\pi$(?) | $\delta_s(CH_3)$ |
| 1369 | w | $\pi$ | $\gamma_w(CH_2)$ |
| 1353 | w | $\pi$ | $\gamma_w(CH_2)$ amorphous regions |
| 1303 | w | $\pi$ | $\gamma_w(CH_2)$ amorphous regions |
| 1170 | vvw | $\pi$ | |
| 1150 | vvw | $\pi$ | |
| 1110 | vvw | $\pi$ | |
| 1080 | vw | $\sigma$ | $\nu$(CC) amorphous regions |
| 1065 | vw(sh) | $\sigma$ | $\nu$(CC) |
| 965 | vvw | ··· | |
| 888 | vw | $\sigma$ | $\gamma_r(CH_3)$ |
| 731 | m | $\sigma$ | $\gamma_r(CH_2)$ |
| 721 | s | $\sigma$ | |
| 600 | vw | ··· | |
| 543 | w | ··· | |
| 200 | vw | ··· | |

[a] $\nu$=stretching, $\delta$=bending, $\gamma_w$=wagging, $\gamma_r$=rocking.
[b] Observed in grating spectra of polyethylene.[9]
[c] Split in *n*-paraffin spectra, with components at 2924 $cm^{-1}$ and 2899 $cm^{-1}$.
[d] Split in *n*-paraffin spectra, with components at 2857 $cm^{-1}$ and 2850 $cm^{-1}$.

**Figure 5.2:** Infrared spectrum and assignments for polyethylene. Reproduced from [12].

**Pattern matching** involves comparing the recorded spectrum against those in a hard-copy commercial or in-house reference library. The result can distinguish polypropylene (PP) from polybutene or poly(ethylene terephthalate) (PET) from poly(butylene terephthalate) (PBT), or indicating the approximate composition of a copolymer [9]. Computerised databases of reference spectra are available with commercial IR spectroscopy equipment companies. The pattern marching has become a more or less standard procedure to conduct directly after collect-

ing an unknown material spectrum.

**Chemical and physical characteristics** can be achieved through an IR spectroscopy study. A higher level of qualitative IR spectroscopy analysis may be considered as correlating to specific spectral features with particular molecular conformations [13], phases [9], orientation [14] or even hydrogen bonding [15]. A most representative example of IR spectroscopy based polymer chemical characteristic is the temperature dependence studies of the $\nu$NH region of polyamides conducted by Schroeder and Cooper [15]. An infrared thermal analysis technique was used to measure the enthalpy of dissociation of hydrogen bonds in various polyamides and to test the influence of the type of nylon, annealing time, and moisture content on the strength of hydrogen bonding. Nearly all of the NH groups were found to be hydrogen bonded at room temperature for every nylon tested. The results show that while the enthalpy of dissociation of the hydrogen bonds was not significantly changed by the type of nylon or any of the tested variables, the modulus was strongly affected by several of the treatments, which indicated that hydrogen bonding does not play a primary role in determining the mechanical properties of polyamides.

### 5.1.1 Ratio method for quantitative IR spectroscopy analysis

The quantitative analysis of FTIR spectra has been highly reliant on the ratio method [16]. It has been widely used and generalised on IR spectra data analysis [10, 11]. In the ratio method, model compounds spectra are needed for calibration [16]. This causes limited application of this method since, it is, in some cases, difficult to get the spectrum of a pure component. Polymer samples have

a relatively simpler molecule structure, and a big body of literature available on their IR characterisation. The standard spectrum of a certain polymer type can be obtained easily, either from the considerable amount of IR spectroscopy polymer studies, or from the build-in spectra library of an FTIR spectroscopy commercial software. The ratio method can be applied to conduct quantitative analysis of plastic IR spectra. The method is introduced in detail as following:

The IR spectrum of a two-component mixture can be represented by:

$$M(v) = f_1(v) + f_2(v) \tag{5.1.1}$$

where $M(v)$ represent the spectrum of a mixture components, $f_1(v)$ is the spectrum of pure component 1, $f_2(v)$ is the spectrum of pure component 2.

The spectrum of a different proportions can be represented as:

$$M(v) = a_1 f_1(v) + a_2 f_2(v) \tag{5.1.2}$$

where $a_1$ and $a_2$ represent the proportion parameters.

The ratio spectrum is given:

$$R(v) = \frac{a_1 f_1(v) + a_2 f_2(v)}{f_1(v) + f_2(v)} \tag{5.1.3}$$

$R(v)$ can be used to define the proportions. In a fingerprint spectral region for component 1, where $f_1(v) >> f_2(v)$, $R(v) \approx a_1$. Conversely, if a fingerprint spectral region $f_2(v) >> f_1(v)$, $R(v) \approx a_2$. This ratio method for proportion determination is not accurate when the band overlap exists or frequency shifts

occur with the different concentrations in the mixtures [16].

### 5.1.2 Polymer blending and copolymer composition

Polymer blends/composites represent a class of materials of great industrial importance, in which two or more materials are combined in a product whose performance/cost ratio is superior to that of the individual components [17]. The polymer-polymer systems are sometimes considered as 'blends' and sometimes as 'composites', and the distinction is not always obvious [17, 18]. Numerous research into the application of IR spectroscopy on polymer blending and copolymer composition study can be found, which, not only proved the great potential of applying FTIR spectroscopy on recycled plastic sorting, but also provide solid support for recycling application of FTIR spectroscopy. Research into the general theory of the interpretation of the infrared spectra of high polymers has a long history [12] and FTIR spectroscopy has been widely applied to provide qualitative analysis of polymer from starting material to finished products [19–23]. The use of vibrational spectroscopy to characterise polymer composites is so widespread that it is impossible to include an exhaustive review herein. Detailed reviews concerning theory, experimental techniques [24, 25] and application examples [17] can be found.

The infrared spectrum of a blend of two incompatible homopolymers will be equivalent essentially to the summation in appropriate proportion of the individual infrared spectra of the two components. Intermolecular interactions, such as hydrogen bonding or polar coupling, may be detected through band position shifts and intensity changes. Pehlert et al. [22] recorded infrared ab-

sorbance spectra at room temperature in the carbonyl stretching region of films of pure ethylene-co-vinyl acetate (EVA) and poly(2,6-dialkyl-4-vinyl phenol) (PDMVPh) blends containing 80, 60, 40 and 20 wt% EVA. The results indicated that a lower wavenumber band is a consequence of hydrogen bonding between the blended material.

Studies of reinforcement and interphases represent one of the most important use of vibrational spectroscopy, and carbon fibre reinforcements is one of the most studied [9, 17, 26, 27]. IR spectroscopy is generally used to detect the presence of different types of carbonyl groups, as well as phenol or alcohol groups to study the surface treatment of CF and its influence on matrix-CF interface [28–32]. Sellitti et al. [29] studied the oxidised rayon-based graphitised carbon fibre using Fourier transform infrared attenuated total reflection spectroscopy (FTIR-ATR). The spectra of the samples oxidised for different times (In Figure 5.3) were compared from the bands at 1720 $cm^{-1}$ and 1580 $cm^{-1}$, arising from the oxidative treatment, followed as a function of oxidation time. Zhang et al. [33] reported the examination of amino-functionalised graphene oxide (GO-$NH_2$) grafted carbon fibre (as shown in Figure 5.4) using FTIR spectroscopy. Compared with untreated carbon fiber, the new feature at 2930 $cm^{-1}$ was assigned as the stretching vibration of -$CH_2$ and the new peaks at 3440 $cm^{-1}$ and 1540 $cm^{-1}$ were related to the N-H stretching modes of amino groups (-$NH_2$). The new features at 1640 $cm^{-1}$ and 1540 $cm^{-1}$ were attributed to amide I (-CO-NH-) and II (-CO-$NH_2$). All the results indicated that partial amino groups of GO-$NH_2$ were reacted with the acyl chloride groups on the carbon fiber surface.

Polymeric fibres are another important class of reinforcement, common types

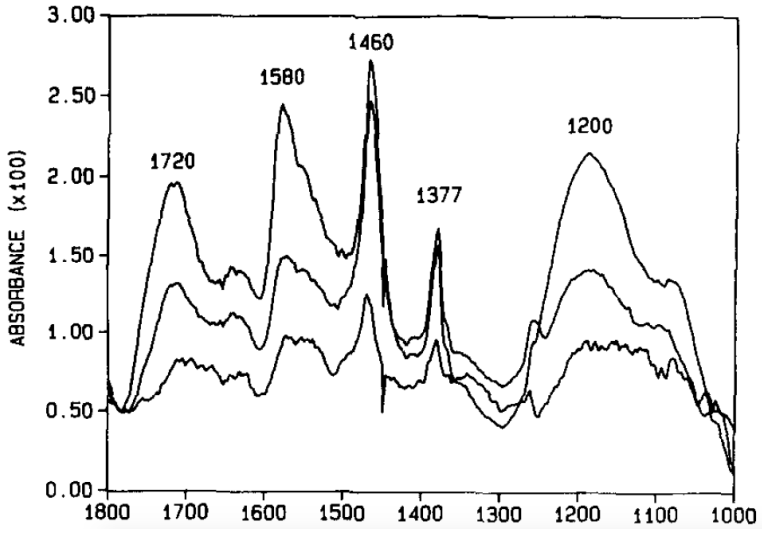



**Figure 5.3:** Spectra of graphitised carbon cloth oxidised for 25 hours (top spectrum), 15 hours (middle spectrum), and 5 hours (bottom spectrum). Reproduced from [29].

being aromatic polyamide or aramid, PET, PE, and PP (polypropylene) [17]. A PE fibres IR spectroscopy study conducted by Van Mele and Verdonck [34] indicate that moisture on the fibre surface interacted with the anhydride hardener and reduced the degree of cure (vulcanisation of rubber) close to the fibre. The interphase in polyethylene fibre/epoxy matrix composites was studied with FTIR microspectroscopy using a set-up to investigate the matrix as close to the fibre as a few $\mu$m or less. Moisture present on the fibre surface was able to influence the polymerisation reaction of the epoxy/anhydride matrix in an irreversible manner. This effect is enhanced for composites from the more hydrophilic polyvinylalcohol fibre. IR spectroscopy was also used to follow the growth of surface carbonyl groups on oxyfluorinated PP fibres [35] as well as the chemical grafting of poly(aniline) and poly(*o*-toluidine) [36] or a diazide derivative [37] to PET fibres.

Particulate fillers are usually added to reduce cost. The most common fillers are

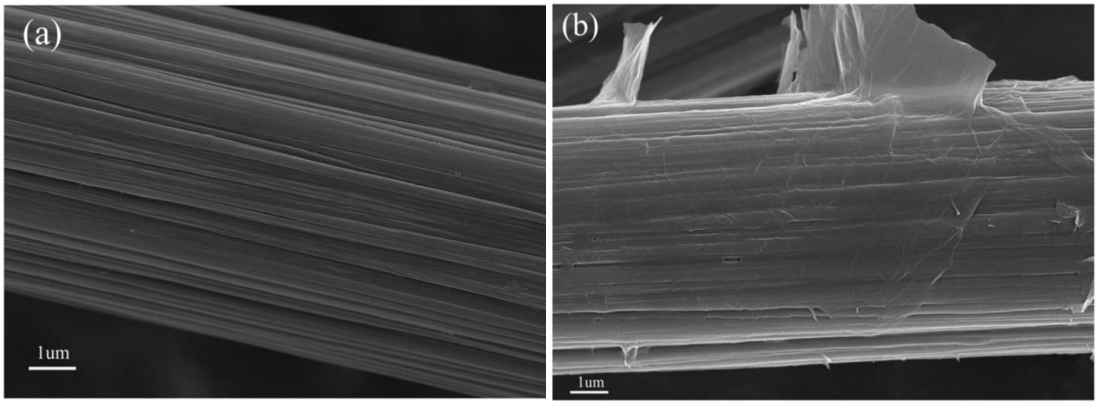



**Figure 5.4:** Scanning electron microscope images of (a) untreated and (b) CF-GO carbon fiber. Reproduced from [33].

silica and silica minerals, calcium carbonate, and metal oxides [17]. Durcova et al. [38] reported quantifying the effectiveness of calcium carbonate dispersion in PP composite fibres using transmission IR spectroscopy. Spectra of composite fibres were recorded through KBr pellets and the absorbance ratio at 714 $cm^{-1}$ and 2721 $cm^{-1}$ was calculated. This ratio was found to be sensitive to the particle size of the calcium carbonate (calcite) present in the polypropylene fibres. The results have shown that FTIR spectroscopy may be reliably applied for evaluation of the relative degree of dispersion of calcite in composite polypropylene fibres in the calcite concentration range 0-15% by weight. The authors believe the FTIR ratio method can be applied to other kinds of polymer/filler composite fibres under the conditions where the following assumptions are valid: (1) the spectrum of the composite fibres is the sum of spectra of individual components; (2) Lambert-Beer's law may be applied; (3) corresponding filler analytical bands are present in the spectra of composite fibres.

Copolymers may be thought of as polymers with chain structures in which

chemically different monomer units are linked in different concentrations in a 'non-uniform' manner [39]. Chalmers [40] reported the infrared spectroscopy analysis of solution (1,1,2,2-tetrachloroethane) cast thin films onto KBr polished plates of poly(aryl ether sulfone) (PES), and two aryl ether sulfone/aryl ether sulfone copolymers. Results have shown that copolymer composition can be determined from a measurement of the absorbance ratio, $1190\,\mathrm{cm}^{-1}/1010\,\mathrm{cm}^{-1}$.

Application of FTIR spectroscopy on polymer blending and copolymer composition study represents a particularly good example of the data processing capability for elucidation of structure information [16]. In this section, we discussed the four major sub-topic in this research field, reinforcement and interphases, polymeric fibres, fillers and copolymers studies, with one of the most representative example given for each topic. FTIR spectroscopy is a powerful tool for polymer analysis. Applying FTIR spectroscopy on plastic recycling has great potential.

### 5.1.3 Application of FTIR imaging to plastic recycling

The use of FTIR imaging on plastic analysis has not been fully explored [3, 41]. ATR-FTIR mapping technique was applied to examine the coarse microscopic two-phase structure of virgin and recycled Acrylonitrile Butadiene Styrene (ABS), High Impact Polystyrene (HIPS) and polycarbonate. The method was compared to optical microscopy both before and after etching with chromic acid. It was found that while optical microscopy showed up phase separation on a scale of tens of microns, the FTIR mapping allowed compositional details to be investigated. The results indicate that, with a spatial resolution of a few

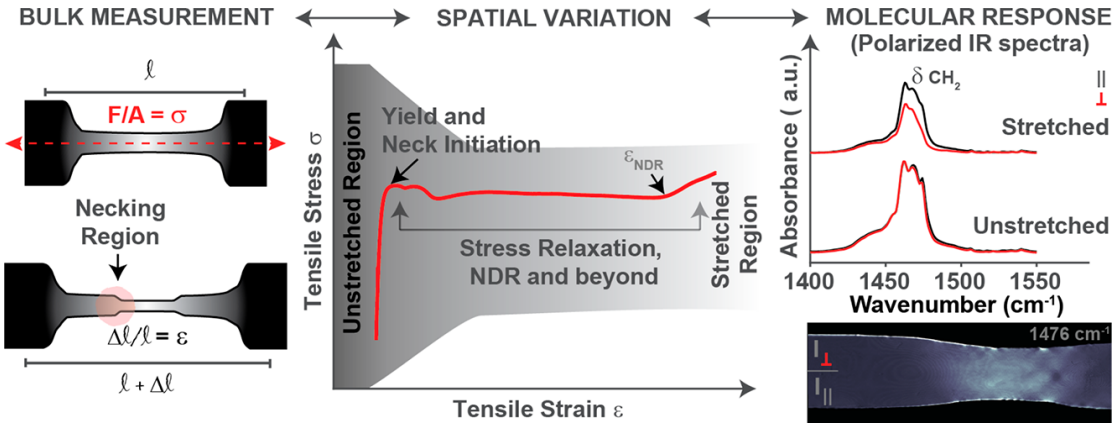



**Figure 5.5:** Schematic illustration of how the spatial variation of molecular response in a polymer can be used to evaluate mechanical properties spectroscopy and imaging. (Left) The physical process of necking, demonstrating the focus of the study on the necking region. (Middle) A representative stress-strain curve for HDPE overlaid on the cartoon of a necked polymer. (Right) Drawn tensile samples exhibit optical anisotropy beyond the yield point that arises from molecular reorganisation and are seen in polarised infrared spectra. Anisotropy measurements by imaging allows visualisation of the spatial distribution of molecular orientation in the entire sample, with localised sensitivity. Reproduced from [41].

microns, FTIR mapping cannot only shows the structure but also gives information on the composition differences between phases [3].

Mukherjee et al. performed the study of applying FTIR imaging to optically and nondestructively measure molecular structure and its spatial dependence in tensile specimens in high density polyethylene (HDPE) homopolymers [41]. Figure 5.5 is the schematic illustration of the methodology. Crystal structure and orientation were obtained from spatially varying measurements of molecular properties in the necking region. Local molecular (re)arrangements to characterise mechanical properties of drawn samples were deduced from spec-

tral data. The study shows that together with multi-scale measurements and analysis, infrared spectroscopic imaging permits a correlation of mechanical properties with its molecular origins. A non-destructively way of examining properties of semi-crystalline polymers was provided.

Compared with ATR-FTIR or single point transmission FTIR, Transmission imaging FTIR has higher requirements in sample preparation, and it takes longer time to collect the data. However, it can provide information that other analytical techniques cannot. The benefits of applying Transmission imaging mode of FTIR spectroscopy to plastic component analyses can be summarised as following:

- To provide a robust form of data. FTIR imaging offers a particular combination of spatial, spectral, and chemical detail [42]. Not only, it reveals information that is available in classical methods, e.g. Thermogravimetric Analysis, but also it provides superior knowledge, e.g. spatial variation of composition by IR spectroscopy imaging or mapping technique.

- To maintain easy lab-industry transformation. Compared with ATR-FTIR, transmission imaging mode relies less on human interaction (ATR requires sufficient sample-crystal contact) in the data collection step. Thus, it has more potential to be applied on a continuous, automatic FTIR plastic examining system, which can be used in an industry setting.

- To provide reliable data. ATR, transflection and reflection FTIR are effectively surface techniques. If the surface of the examined point is not representative of the whole plastic bulk, for example, due to plasticiser mi-

gration to the surface [43], the spectrum obtained from those FTIR modes can be misleading.

For plastic recycling, the physical or chemical properties of plastics can be manipulated by copolymers or blends and many of the waster plastics are copolymers or blends and this is problematic as the composition information is often missing when it comes to the recycling stage. The quality of recycling plastic products highly depends on the quality of the sorting process and thus knowledge of the composition of the polymer and polymer blend is key in order to facilitate effective and efficient reuse of the plastic material. Studies [44, 45] have indicated the composition was more significant influence other than whether the materials were recycled or not. To produce better quality recycled plastic for advanced technical products, it is essential to understand the precise components of the input material, to adjust blends and processing conditions accordingly.

## 5.2 Aims

The application of recycled polymers and composites is generally limited to lower end applications, in part because the finely tuned properties required in high-end products can-not be guaranteed from recycled materials. This is because recycled plastics come from a wide variety of sources, with different compositions and conditions. As a result, different reprocessing blends and parameters are needed for each batch of recycled plastic to achieve a product with consistent properties.

This research discusses the application of FTIR spectroscopy to recycled composite production, focussing on studies in the following areas:

1. Analysis the effects of pre-processing. The pre-processing step in FTIR imaging data analysis is a very delicate procedure. With the abundant literature on polymer IR spectra, Section 6.1 is a knowledge-based approach study of pre-processing methods for IR spectrum analysis. Different methods were compared and the best performing parameters were tuned to provide the best S/N ratio of the spectra.

2. Examine the ways in which FTIR imaging can be used to the component study of VPP + Talc plastic sample. The identification and characterisation of the individual components present in the composite using FTIR imaging are provided in Section 6.2. In this case study, the building of a spectra library, the calibration of talc concentration with IR spectrum peak ratio/integration, and the application of the calibration result to analyses the un-even distribution talc in a film sample were discussed in detail.

3. Investigate the prominent benefits of applying FTIR imaging technique to investigate the composite structures of 'real-world' composites sample for recycled industry plastics, including: ABS + $CaCO_3$, PP + talc, PP + rSMCF + SF, PP + MAPP + rSMCF + SF and PP + PET morphology on the micron scale. Emphasising on the interaction between different materials, the interface interaction and the distribution of each composite were studied in detail.

The purpose of this study is to investigate the use of FTIR spectroscopy as an

analytical tool for process control during the plastic recycling and to use high-definition FTIR imaging to monitor the chemical composition and spatial distribution of plastic composites. We hope this study can lead to the possible application of FTIR spectroscopy as online analysis methods for plastic recycling.

Plastic sample pellets with given composition were prepared by Mr Bing Wang. Details of the plastic films preparation and FTIR imaging are detailed described in the Experimental Section. The resulting FTIR imaging data and analysis are reported in the following Chapter.

CHAPTER 6

# Plastic samples analysis using FTIR spectroscopy

## 6.1 Studying the influence of pre-processing steps

In order to improve data quality, pre-processing is normally the first step taken in FTIR spectroscopy data analysis. Based on the purpose of the step, it can be divided into: spectra correction, de-noising, normalisation and mathematical derivatives. Two or three methods are often combined, and different options are available for each of the pre-processing steps. In this section, with the clear structured virgin polypropylene (VPP) plastic film sample, a **data-processing** $\rightarrow$ **result** $\rightarrow$ **data-processing** loop can be developed to gain better understand of the data pre-processing steps and algorithms used in each step. The PP material used in this study is an injection grade PP block copolymer produced by Dushanzi Sinopec, denoted EPS30R.

Typical pre-processing steps for plastic FTIR data is as following:

1. Spectra are truncated to exclude regions with little or low information;
2. Baseline correction is applied to correct any baseline offset;
3. Smoothing is applied to reduce the noise level of the spectrum;
4. Normalisation is applied to correct sample thickness's influence on the spectra;

### 6.1.1 Spectra correction

Spectra correction includes the correction of the baseline caused by scattering and spectral contributions that arise from atmospheric water vapour, carbon dioxide, paraffin or other interfering compounds. For bio-material, scattering is one of the main causing for spectra distortion, extended multiplicative scattering correction (EMSC) [46] is normally applied. For plastic samples, the scattering is less problematic. In this section, baseline correction is achieved using the 'msbackadj' function within MATLAB, in which a spline approximation regression is used to adjust the baseline. The approximation is obtained using multiple shifted windows across the x-axis. Two parameters, 'WINDOWSIZE' and 'STEPSIZE' need to be tuned to get the best performance. 'WINDOWSIZE' sets the width for the shifting window. For instance, 'WINDOWSIZE' = 200, means a background point is estimated for windows of 200 separation unit. 'STEPSIZE' sets the steps for the shifting window. 'STEPSIZE = 200' means an estimated for windows at every 200 separation unit. In the case of IR spectrum, the separation unit wavenumber is 1 $cm^{-1}$.

Figure 6.1 is the spectrum truncation and baseline correction on the mean spectrum of 36,864 (six-by-six imaging using 32 $\times$ 32 Focal Plane Array detector)

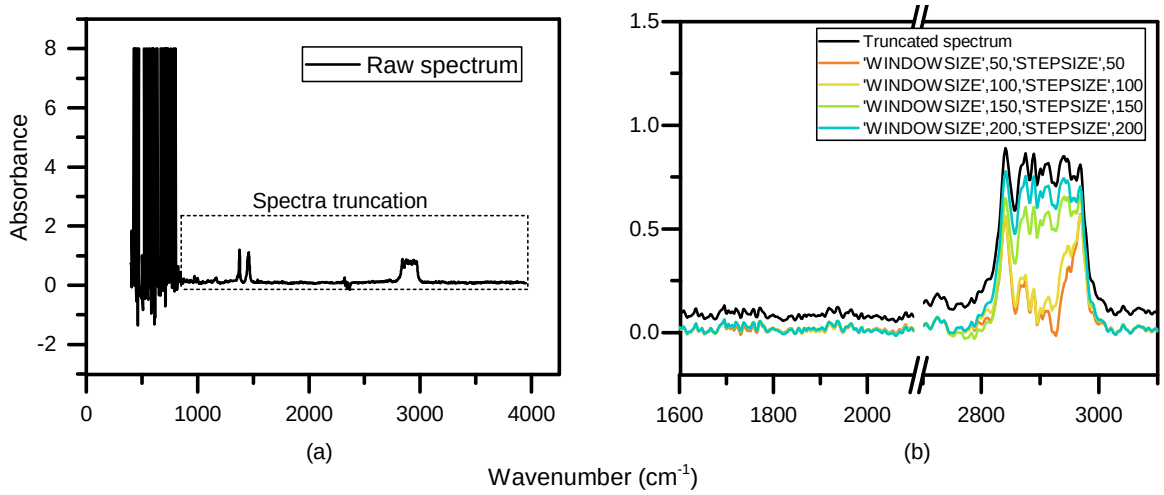



**Figure 6.1:** PP spectrum truncation and baseline correction. (a) is the raw spectrum and the selected wavenumber range for spectrum truncation step; (b) is the truncated spectrum and the applying of 'msbackadj' function on different parameter setting. Wavenumber range 1600 - 3000 $cm^{-1}$ is shown in (b) to better visualise spectral changes.

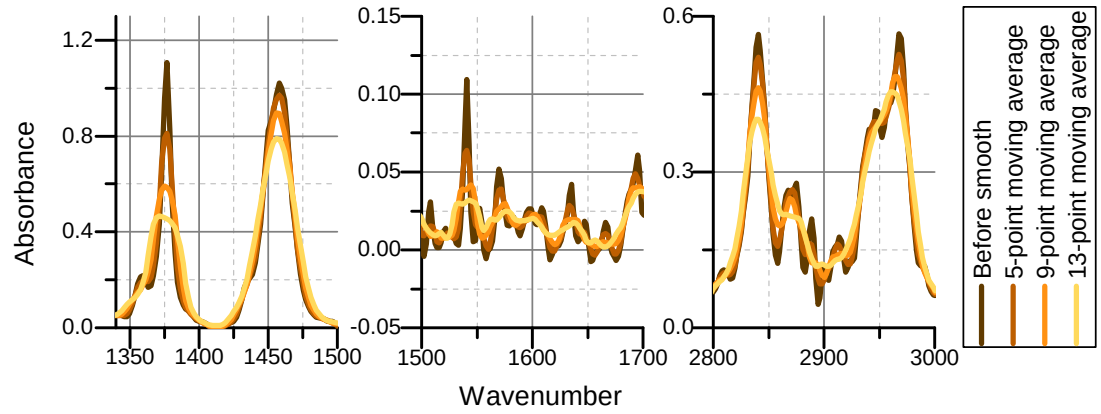
VPP IR absorption spectra. Absorbance in the raw spectrum 450 - 850 $cm^{-1}$ wavenumber range have no IR transmission. Excluding that wavenumber region to reduce the data size thus increase the computational efficiency is necessary. Figure 6.1a shows the raw spectrum and the selected wavenumber range 850 to 4000 $cm^{-1}$ which is highlighted inside the rectangular box. Figure 6.1b is the truncated spectrum and the application of 'msbackadj' function using different parameter settings. Comparing the truncated spectrum (black line in Figure 6.1b) with the resulting spectra (orange, yellow, green and blue line in Figure 6.1b), the baseline region of the PP spectrum (1600-2000 $cm^{-1}$) is moved to zero absobance, indicating that 'msbackadj' function is effective in correcting the baseline problem of plastic IR spectra. In the 2800-3100 $cm^{-1}$ region, more spectrum distortion can be observed by the 'msbackadj' with lower 'WIN-

DOWSIZE' and 'STEPSIZE' value, indicating the different performance of the algorithm with different parameter setting. 'WINDOWSIZE' 100 and 'STEPSIZE' 100 gives the closest result with PP spectra reported in the literature [47–49].

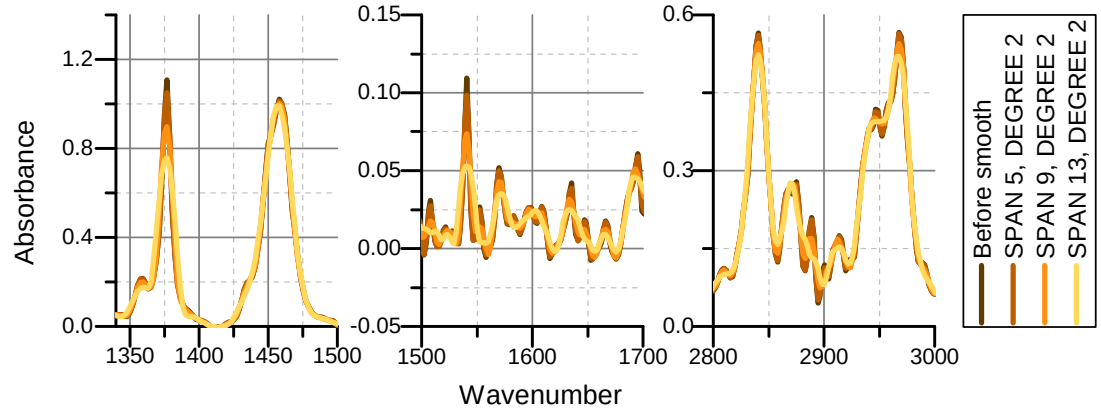
### 6.1.2 De-noising

De-noising is also called smoothing, aiming to improve the signal-to-noise ratio of the data. Savitzky-Golay (S-G) [50] smoothing is the most applied method for FTIR spectroscopy data sets [51, 52]. By fitting successive adjacent data points with a low-degree polynomial using the method of linear least square, the S-G method has good performance in smoothing data with shape peaks. Two parameters, 'SPAN' and 'DEGREE' need to be chosen to get the best performance of the algorithm. 'SPAN' defines the number of adjacent data points used to compute each element of the de-noised data. 'DEGREE' specifies the degree of the polynomial to be used in the S-G method. Noting S-G smoothing could introduce distortions in the spectral measurements, and at the same time smooth out important information in the spectra, S-G algorithm parameters need to be carefully chosen.

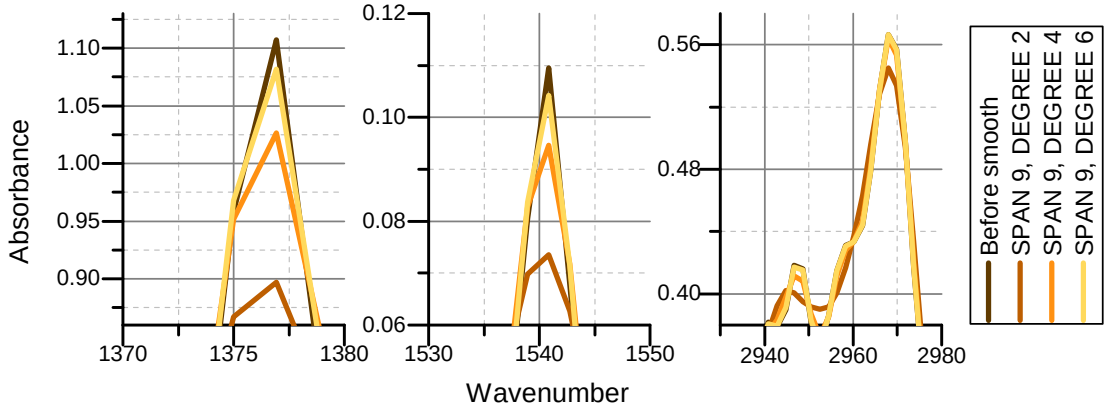
Moving average is another commonly used smoothing technique in signal processing. By creating a series of averages of different subsets of the full data set, it aims to smooth out fluctuations is not the main trends of the signal. Given a fixed subset size, the first element of the moving average is obtained by taking the average of the initial subset. Then the subset is 'shifting forward' to obtain the next moving average value, until it reaches the end of the data set.




**(a)**




**(b)**




**(c)**

**Figure 6.2:** Comparison of moving average and S-G smooth method. (a) n-steps moving average smoothing under the SPAN value of 5, 9 and 13; (b) Savitzky-Golay smoothing using two-degree polynomial fitting under the SPAN value of 5, 9 and 13; (c) Savitzky-Golay smoothing using different degree (2, 4 and 6) polynomial fitting under the SPAN value 9.

Figure 6.2 is the application of n-steps averaging smoothing and S-G method on PP spectrum (after truncation and baseline correction). Figure 6.2a is the comparison of 5, 9 and 13 points of moving average smoothing with the original spectrum. The corresponding wavenumber length for 5, 9 and 13 points are 9.64, 17.36 and 25.04 $cm^{-1}$ respectively. Zoom-in images of wavenumber region 1340-1500 (shape peak), 1500-1700 (baseline) and 2800-3000 (peak) $cm^{-1}$ are provided. It can be observed that the spectra distortion increase as the 'n' value increase, meanwhile the algorithm's ability to smooth out the noise decreases. In 5-steps averaging smooth case, less peak loss can be observed in the shape peak region, yet the baseline region remains noisy. When increasing the 'n' value to 13, less noise can be observed from the baseline region. Nevertheless, the peak lose in a shape peak region is large.

Applying the S-G method using different SPAN setting was tried. Fixing DEGREE to 2, S-G smooth result with 5, 9 and 13 SPAN value were compared. The result is given in Figure 6.2b. Compared with Figure 6.2a, less peak lose is caused under the same SPAN value. Correspondingly, less de-noise effect can be observed from the baseline region. Compared with moving average results, S-G algorithm can provide better-resolved peaks [53]. Figure 6.2c is the S-G de-noising using different 'DEGREE' of polynomial function. The result indicates that higher degree of the fitting function can cause less peak distortion (provide better-resolved peak). However, it weakens the smoothing algorithm's ability to reduce high-frequency noise.

In general, it is difficult to draw the conclusion which smooth algorithm on what parameter setting works the best, as both de-noise algorithms tested here

decrease its smooth ability when trying to keep the shape peak resolution. For the PP IR spectrum, S-G de-noising algorithm with SPAN = 9, DEGREE = 2 was selected.

### 6.1.3 Normalisation

Normalisation is used to correct the influence of thickness. The different thickness or concentrations of the samples can sometimes cause the most prominent source of spectral variation. It is important to normalise the IR spectroscopy data to minimise the influence of the varying thickness of the sample. Peak maxima normalisation and vector normalisation are the most popular used methods [54, 55]. As with most pre-processing steps, normalisation algorithms need to be carefully selected and applied with caution.

Each IR spectra is corresponding to the chemical information contained in that pixel. In an ideal situation, excluding the influence of machine, environment noise/error and sample preparation variation, spectra collected from the same VPP film should be identical, as the material composition is uniformly distributed across the sample. For 'real-world' data, the spectra from the same VPP film cannot be identical, as noise is inevitable. The noise level, however, can be represented by the distribution of the spectra in the sample.

In statistics, density distribution is normally used to compare between populations (a collection of individual samples). Taking spectra from the same sample as a population, each spectrum as an individual, the probability density distribution of wavenumber can be obtained. In the ideal case described above

(spectra from the population are identical); probability density distribution of one wavenumber absorbance will be a pulse (a straight line in parallel with the y-axis). For real experimental data, the probability density distribution is usually a normal distribution (or close to normal distribution). The higher the S/N ratio is, the sharper the normal distribution will be, as a higher S/N ratio representing a closer to ideal experiment setting. The shape of the normal distribution can be used to evaluate the S/N ratio of the data set. By comparing the S/N ratio level of the data set before and after one pre-processing step, the pre-processing algorithm's performance can be evaluated.

Figure 6.3 is the evaluation of pre-processing steps on PP FTIR imaging data. Figure 6.3a is on single wavenumber absorbance value. Figure 6.3b is on peak wavenumber absorbance integration value. Three pre-processing steps: baseline correction (msbackadj), smooth (S-G), and normalisation (peak value normalisation) is evaluated. Four data set, resulting from the three steps pre-processing: raw data set (blue), data set with baseline correction only (orange), data set with smooth after baseline (yellow), and data set with all three pre-processing steps (purple) are presented.

In the figure, as we expected, all the density distribution is normal or close to a normal distribution. From 'Raw' to 'Baseline', then to 'Smooth', all those steps are improving the data quality as the density distribution is getting shaper and shaper.

The peak maxima normalisation method is used and tested in this section. The normalisation method relies on dividing the absorbance of a certain wavenumber by the absorbance on 1377 $cm^{-1}$. For PP film, the spectra noise resulting

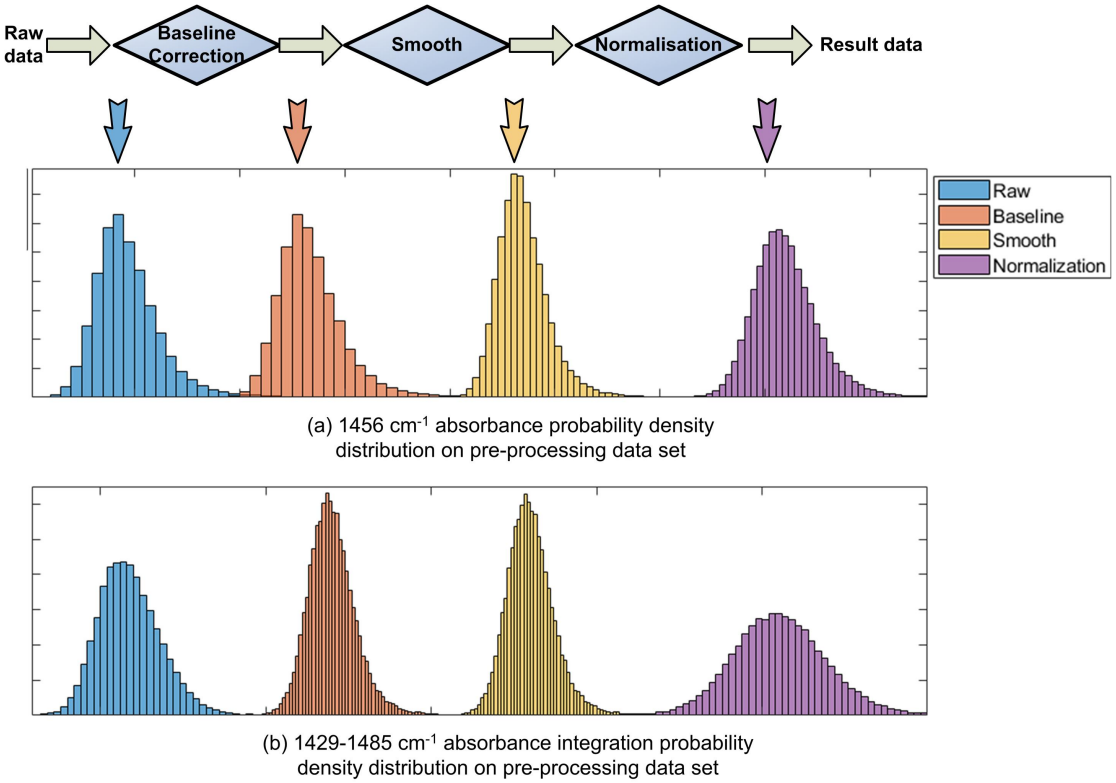



**Figure 6.3:** Evaluation of pre-processing steps based on VPP FTIR imaging data. Probability density distribution of the FTIR imaging data is produced based on data sets resulting from each pre-processing step. The shaper the distribution the better the S/N of the data set. (a) probability density distribution on wavenumber 1456 $cm^{-1}$ absorbance value; (b) probability density distribution on wavenumber 1429-1485 $cm^{-1}$ absorbance value integration. Blue: raw data set; orange: data set after baseline correction; yellow: data set after baseline $\rightarrow$ smooth; purple: data set after baseline $\rightarrow$ smooth $\rightarrow$ normalisation.

from thickness variation is relatively low. The correction of the thickness variation introduced higher noise, as single wavenumber 1377 $cm^{-1}$ absorbance contains a high level of noise. This is confirmed in Figure 6.3. After normalising, the distributions are getting flatter in both (a) and (b). This means the normalisation steps for the PP imaging data is introducing noise other than re-

ducing noise.

The phenomenon observed in Figure 6.3 does not means that normalisation should not be used at all. All normalisation algorithms introduce noise to the data set. However, if the noise caused by the thickness variation is much higher than the noise introduced by normalisation, it is still worth normalising the spectra to correct the thickness influence.

In Figure 6.3, both (a) and (b) indicated that the distribution changes to a sharper shape from raw data set to baseline correction data set (blue to orange). Comparing with the shape change in (a), the shape change in (b) from blue to orange is larger. This is because the density distribution in (b) is based on peak integrations (multiple absorbances) other than one single wavenumber absorbance.

Baseline correction improved little of the spectra quality (the distribution in (a) changed little from blue to orange). It, however, improved the data quality a great deal, in terms of the peak integration value (the distribution in (b) changed a lot from blue to orange). This agreed with that statement that, by combining multiple wavenumber absorbances, the S/N ratio can be increased, which has been drawn by many researchers [56].

### 6.1.4 Mathematical derivatives

Mathematical derivatives can reveal the subtle differences between spectra hidden by overlapping bands. The bands' full width could be narrowed at half height value (FWHH) by applying differentiation to the spectrum. However, each differentiation amplifies noise [57]. Furthermore, it is difficult to link the

data pattern with biochemical or physical knowledge, as the transferred data after mathematical derivatives are not directly resulted from light absorption.

### 6.1.5 Section conclusion

In this section, a good collection of spectrum pre-processing methods were discussed. Based on the VPP sample FTIR imaging data, utilising statistical density distribution as tool, the performances of each method were tested and compared. This information is used in the subsequent sections when preprocessing IR spectroscopic data on polymeric samples.

## 6.2 Case study: Virgin polypropylene with talcum powder

### 6.2.1 Studying PP/Talc composition using ATR-FTIR spectroscopy

PP samples containing 0 to 20 wt.% talc concentration with step size 2.5% were prepared, and the ATR-FTIR spectra recorded. The variation of talc concentration was achieved by changing PP/talc feeding composition ratio in the regular blending and extrusion process, as shown in Figure 6.4. For each talc composition, a minute of the extrusion pellets were collected during the steady period (20 minutes after changing the composition setting) to maintain a low variation between samples in the same group. Five pellets were randomly selected from pellets collected between 20 and 21 minute results, ten ATR-FITR spectra were collected for each pellet. This data was used to calibrate the PP/talc IR spectrum peak ratio against wt.% talc of the composite. This dataset is part of the result of another project studying the plastic extrusion performance; and this

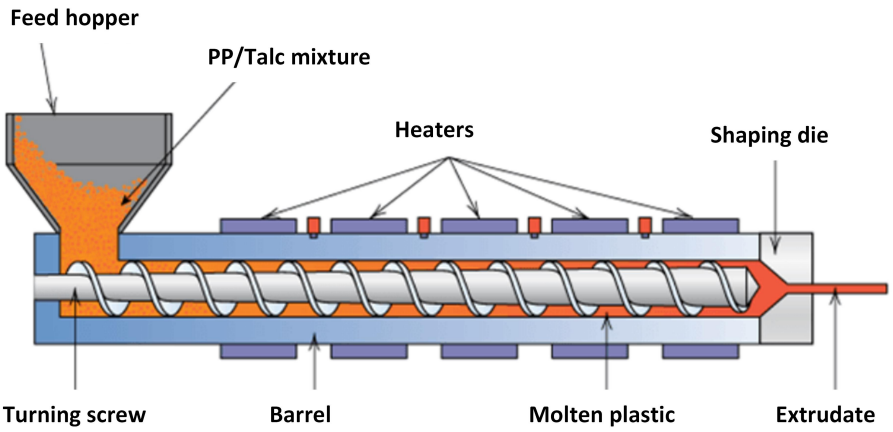



**Figure 6.4:** Overview of making PP/Talc pellets with fixed talc concentration in regular blending and extrusion process. Reproduced from [58].

data collection was conducted by Mr Bin Wang.

The ATR-FTIR spectra were collected using a spectral resolution 4 $cm^{-1}$ and co-adding 16 scans. Each spectrum was normalised at the PP IR absorption peak at 1166 $cm^{-1}$. No other pre-processing methods were used to maintain minimum manipulation of the spectra. The data set was divided into nine groups based on the talc concentration, and each group contained 50 spectra. Statistical analysis, boxplot was used to visualise the inter and intra group relationship of peak ratio between 1020 (talc IR absorption peak) and 2916 $cm^{-1}$ (PP IR absorption peak). By combining multiple wavenumber absorbance, the S/N ratio of the data can be improved. Therefore, the same boxplot was generated based on the talc peak integral (1010-1030 $cm^{-1}$) to compare with the single peak ratio result.

Figure 6.5 is the analysis results of the ATR-FTIR spectra. Both the peak ratio and peak integral value boxplot show that the intra-group spectra variation

gets larger as the concentration of talc increases. This suggests that the mixer has dropped performance on higher talc concentration plastic. Comparing with the boxplot based on peak ratio (Figure 6.5a), peak integration boxplot shows a more steady tendency (based on the mean, median, quartile one, quartile three and the outlier identified) in terms of the intra-group variation increase.

Beer's law states that the concentration of a chemical solution is directly proportional to its absorption of light. The ratio between the talc peak and PP peak should be in a linear relationship with ratio of talc concentration and PP concentration. As the material is composited by PP and talc only, the PP concentration equals to one minus talc concentration. Using the mean value of each group after outlier cut, the fitted 1st degree polynomial (with R-square=0.9948) is given, and the result is plotted in Figure 6.5c:

$$\frac{Concentration_{talc}}{Concentration_{PP}} = 0.3886 \cdot peakratio - 0.0386 \tag{6.2.1}$$

As all the spectra were normalised to PP peak (1166 $cm^{-1}$) before the integration value were obtained. It is a linear relationship between the talc peak integral and ratio of talc concentration and PP concentration, using the mean value of each group after outlier cut, the fitted 1st degree polynomial (with R-square=0.9866) is given, and the result is plotted in Figure 6.5d:

$$\frac{Concentration_{talc}}{Concentration_{PP}} = 0.00597 \cdot peakintegral - 0.04988 \tag{6.2.2}$$

In both Figure 6.5c and Figure 6.5d, the experiment value fitted well with the linear relationship prediction. The linear relationship obtained in this result

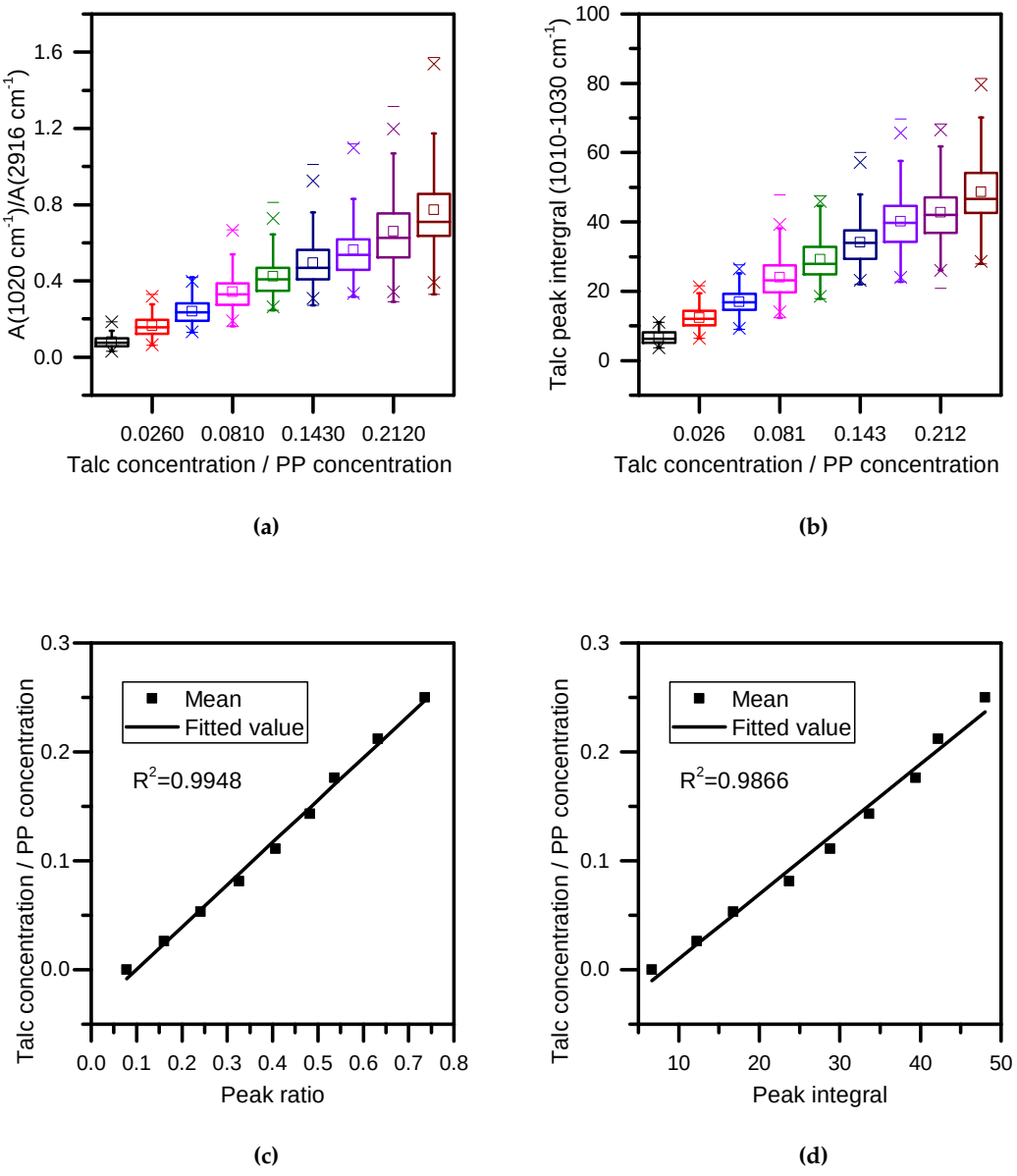



**Figure 6.5:** PP/Talc composition ratio IR spectroscopy study. (a) boxplot of peak ratio between wavenumber 1020 $cm^{-1}$ and 2916 $cm^{-1}$ of each concentration group; (b) boxplot of talc peak integral (wavenumber 1010 $cm^{-1}$ to 1030 $cm^{-1}$) of each concentration group; (c) scatter of mean peak ratio of each group and the linear fit result; (d) scatter of mean peak integral of each group and the linear fit result.

agrees well with other IR spectroscopy proportion analysis [16, 17, 22]. Even though, in terms of the goodness of fitting ($R^2$), peak ratio result is slightly better. The peak integration calibration result was used to obtain a quantitative analysis of a talc/PP film sample in the next section, due to the better S/N ratio of the data set.

### 6.2.2 Studying talc reinforced PP samples using FTIR imaging

After confirmation a calibration plot for the wt.% talc in PP composites, the samples were investigated further through examination of their microstructure using FTIR imaging. FTIR imaging was performed on a 80% VPP, 20% talc (VPP + 20%talc) sample under the following conditions: pixel size 1.1 $\mu m^2$, spectral resolution 2 $cm^{-1}$ and co-adding 16 scans. Using the film making procedure developed in the Experimental Section (see page 300), the thickness of the result sample film varies from 20 to 25 $\mu$m across the sample (the lowest and the highest value of those five measurements).

'msbackadj' function (WINDOWSIZE 60, STEPSIZE 60) baseline correction and nine point moving average smooth algorithm were applied before each spectrum is normalised at the PP IR absorption peak at 1166 $cm^{-1}$.

The pseudo-colour image based the integral band area of the talc (1010-1030 $cm^{-1}$) is shown Figure 6.6c, where outliers (pixels in the high and low range) were cut off. *Quartile one* - (1.5 x *inter quartile range*) and *quartile three* + (1.5 x *inter quartile range*) were used as the boundary. Using the 20, 40, 60 and 80 percentiles of the values as boundaries, the selected pixels were grouped into

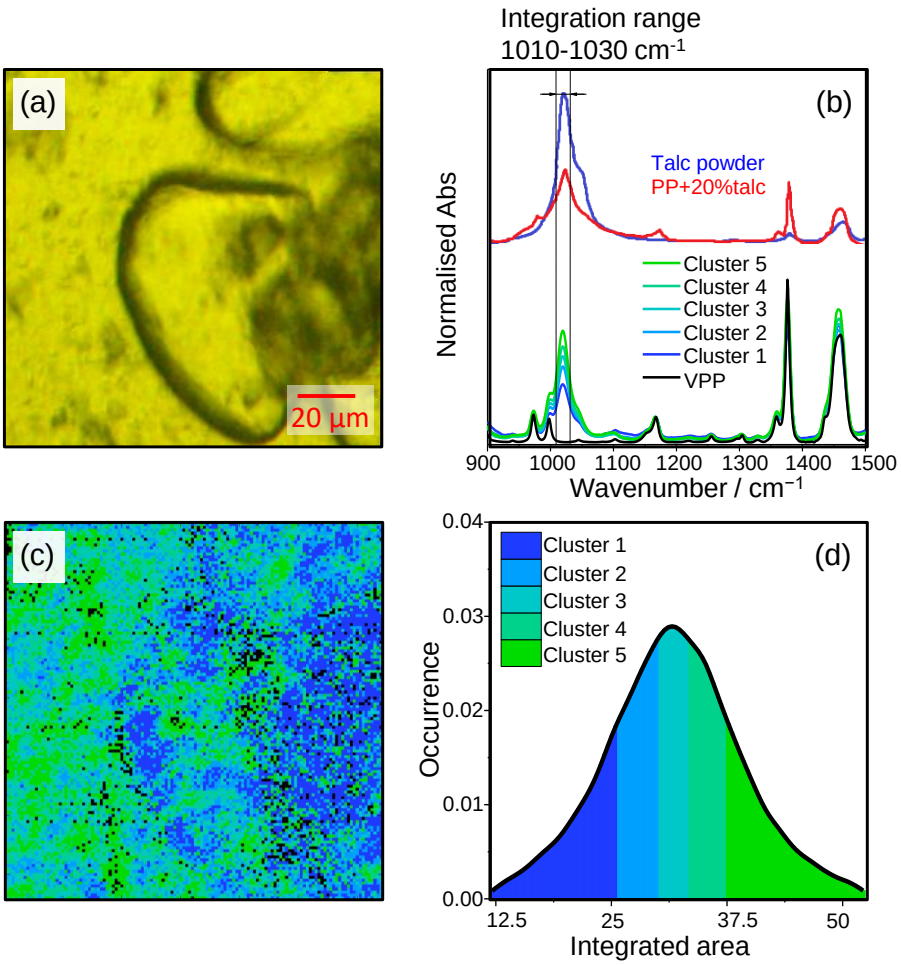



**Figure 6.6:** Images of VPP + 20%talc sample. (a) Optical image of the selected area for IR spectroscopy examination. (b) Medium spectrum of each cluster, the standard spectrum of PP, Talc powder and PP + 20% talc. (c) Pseudo-colour image based on FTIR imaging results. Each block represents a spectrum acquired at this position. Pixels of the same colour indicate the spectra at those position falls within the same cluster. (d) Integral band area 1010-1030 $cm^{-1}$ absorbance histogram. The dark region observed in the optical image contains less talc compared with the bright region.

five clusters. Five different colours were given to pixels in different clusters.

From the image, we can see the sample has higher talc percentage in the right than the left. The talc is not evenly distributed throughout the sample. A dark region can be observed in the right half sample in the optical image. The shape of the low talc region in the pseudo-colour image matches well with the shape of the dark region in the optical image.

The standard transmission spectrum of talc powder and PP with 20% talc are given in Figure 6.6b. The top half of the figure is the comparison of the standard talc spectrum with standard VPP + 20%talc sample spectrum. The bottom half figure is the comparison of the mean spectrum of each cluster with standard VPP spectrum. The integration range is indicated. In the figure, the integration range 1010-1030 $cm^{-1}$ is the wavenumber range that PP has no IR absorption, while talc has its main fingerprint absorbance peak. In a low talc composition cluster, its mean spectrum is closer to standard VPP + 20%talc sample spectrum. In a high talc composition cluster, its mean spectrum is closer to standard talc spectrum. The mean spectra change from being closer to VPP + 20%talc to be closer to talc spectrum. This indicated the major difference between clusters is talc composition.

As shown in Figure 6.6d, the talc peak integral value increase from cluster one to cluster five. The mean value for each cluster is: 21.4816, 27.9932, 31.7078, 35.3946 and 42.0204. Using the peak integral to talc concentration fit result (Equation 6.2.2 on page 123), the mean talc concentration for cluster one to five are: 7.27%, 10.49%, 12.24%, 13.90% and 16.73% respectively. This result is a bit lower than 20% weight ratio for talc/PP feeding rate. This may due

to the imaged region has low representative of the whole sample, or the low performance of the mixer in the extrusion process.

Further explore the reason for the dark region and the low talc percentage in the region by applying two-cluster Fuzzy C-Mean Clustering on the whole spectrum of the raw data set. The pseudo-colour image and the medium spectrum of the cluster are given in Figure 6.7a and b respectively. This approach easily distinguished the dark region from rest of the PP/talc matrix. The medium spectrum of the cluster indicates that spectra from cluster two (dark region) have higher baselines. A consistent baseline level, which is normally the result of over thickness, can be observed all over the medium spectrum of cluster two.

Baseline problem in FTIR spectroscopy data can normally be the result of over thickness or scatter. We think that the dark region is the result of poor mixing in the extrusion step. A piece of VPP is not melted and mixed with talc properly in the entrusted pellet, which is further causing an over-thickness problem in the film making process. Possibly, the dark region in the optical image, and the high baseline region in the FTIR image are due to the over-thickness.

### 6.2.3 Section conclusion

Talc reinforced PP is one of the most widely used plastics in automobile industry. The recycling of this plastic blend is problematic as different manufacturer use various talc composition. The high-end reuse of the blend requires a precise evaluation of the talc composition. We presented a accurate quantitative calibration that can link the IR absorption spectrum with the talc weight

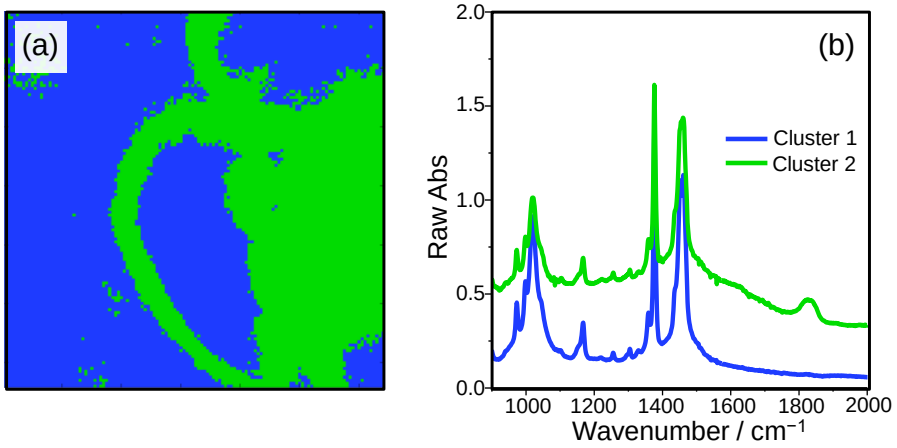



**Figure 6.7:** Images of VPP + 20%talc sample. (a) Pseudo-colour image based on raw data on the whole spectrum. (b) Medium spectra of each cluster, which shows the major difference between the clusters is the spectrum baseline.

percentage, based on ATR-FTIR data. Micron scale FTIR imaging was conducted on talc/PP film sample. pseudo-colour image visualising the distribution of talc in the PP matrix indicated a highly uneven distribution. Utilising the talc ratio & IR peak ratio calibration, the uneven distribution was quantified, which provided an accurate composition evaluation of talc/PP plastic. The great potential of FTIR on composition study of recycled plastic is revealed. The methodology, building a spectra library $\rightarrow$ calibrating composition & IR absorption $\rightarrow$ quantitative composition evaluation, can be applied to other types of recycled plastic.

## 6.3 FTIR imaging for plastic film composition identification

Using reinforcing agents to improve the performance of thermoplastics has been the focus of research in recent years [59–61]. As well as the reinforcing

itself, the percentage composition influences the physical and chemical properties. Studying the agent concentration in recycled reinforced plastics, as well as the the spacial distribution can help develop more precise recycling methodologies, improving recycled plastic's physical and chemical performance. In this section, virgin polypropylene with short milled recycled carbon fibre (VPP + 2%rSMCF + 1%SF), virgin polypropylene, maleic anhydride grafted polypropylene with carbon fibre (VPP + 20%MAPP + 5%rSMCF + 2.5%SF), acrylonitrile butadiene styrene with calcium carbonate (ABS + 20%$CaCO_3$) and virgin polypropylene with poly(ethylene terephthalate) (VPP + 20%PET), were investigated using FTIR imaging.

### 6.3.1 Virgin polypropylene with short milled recycled carbon fibre

The automobile industry has an increasing interest in polypropylene (PP), due to its low lightweight index, short cycle time and great recycling potential [62]. Carbon fibre (CF) as an important reinforcing material for thermoplastic materials has been widely used, in particular CF reinforced PP composites have been extensively studied [49, 63]. It has been widely accepted that the inter-facial properties between CF and PP matrix is critical for the ultimate performance of the resulting composites [47]. In this section, FTIR imaging was applied to CF reinforced PP composites to study the component distribution in the PP matrix. The CF material used in this study is recycled CF (rCF). Details of the recovering process and the characterisation can be found in the paper by Wong et al. [48]. The rCF material was shredded using a cutting mill (Retsch SM2000) with a 0.5 $\times$ 0.5 mm square aperture sieve to obtain the short milled recycled car-

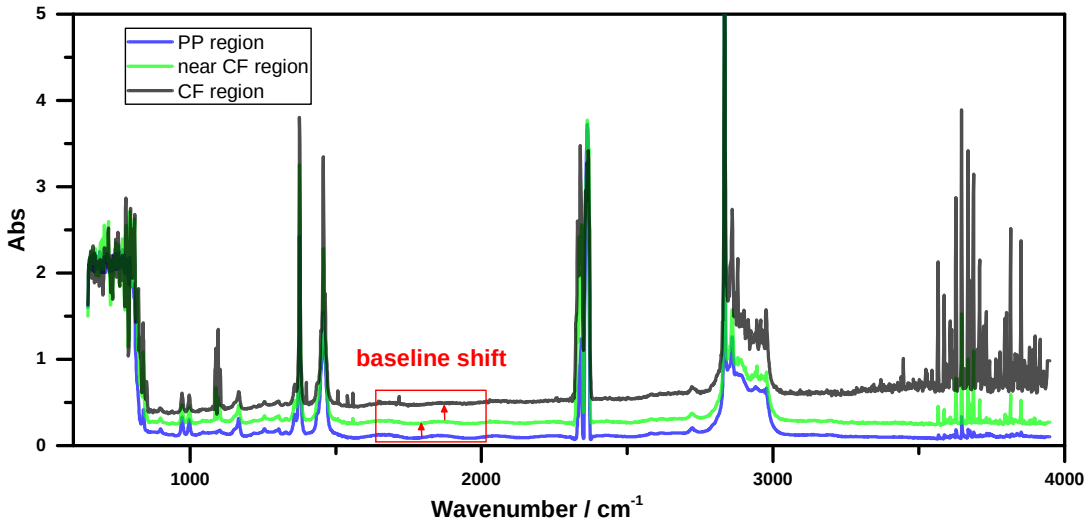
bon fibre (rSMCF). Due to the removal of coatings on the fibre surface during the recycling process, the rSMCF material comes in a fluffy form with severe entanglements [64]. In the sample preparation, rSMCF was dispersed using polyoxyethylene (20)oleyl ether as a surfactant (SF) which was compounded with the polymer matrix via compression, extrusion and injection moulding cycle [64]. Utilising FTIR imaging, inter-facial interaction between fibres and PP matrix was analysed to detail investigate the existence of polyoxyethylene (20)oleyl and its distribution in the final PP composite.

FTIR imaging in the region of interest was performed on 97% VPP, 2% recycled short milled carbon fibre (rSMCF) and 1% surfactant sample (VPP + 2%rSMCF + 1%SF). 128 $\times$ 128 pixels were collected in each sample by co-adding 16 scans at 2 $cm^{-1}$ spectral resolution (wavenumber step size 1 $cm^{-1}$). Transmission in 'High magnification' mode was used, which gives pixel sizes of 1.1 $\mu m^2$.

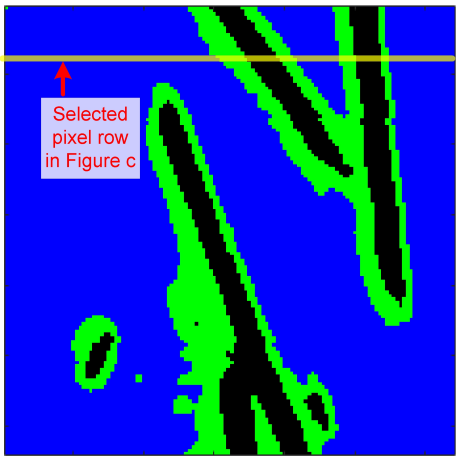
### rSMCF identification

The diameter of the solid cylinder-shaped rSMCF (5-10 $\mu$m) can induce Mie scattering in the transmitted mid-IR light [65, 66]. Mie scattering caused by the micrometre-sized rSMCF can be used to identify their location within the polymer matrix.

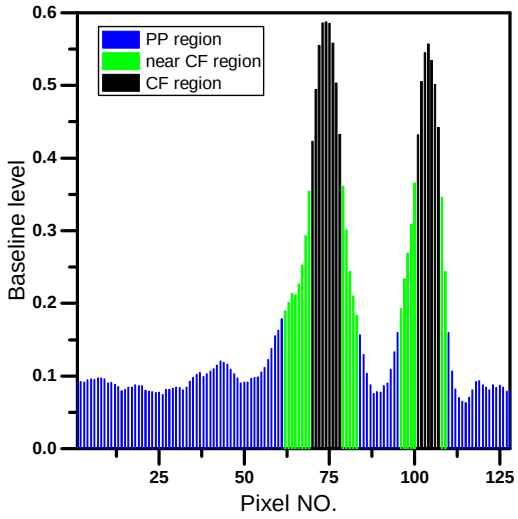
Wavenumber range 1800 $cm^{-1}$ to 2000 $cm^{-1}$ (highlighted in Figure 6.8a), a region in which there should be no peaks due to the surfactant or PP, was selected. Mean absorbance of each spectrum in this baseline range was obtained and used as a feature value representing that spectrum. This mean absorbance




(a)




(b)




(c)

**Figure 6.8:** Identification of rSMCF region based on the baseline shift level. (a) Mean spectrum of clusters. Clear baseline shift can be observed between PP region (blue line), rSMCF nearby region (green line) and rSMCF region (black line). (b) Pseudo-colour image resulting from Fuzzy-C Mean clustering, with PP region (blue), rSMCF nearby region (green) and rSMCF region (black) belongs to three clusters. (c) Selected pixel row, 1800-2000 $cm^{-1}$ wavenumber region mean absorbance bar plot. Compared with PP region (blue), rSMCF region (black) and its nearby region (green) have a higher baseline level.

value in the baseline wavenumber range is called the baseline level. Figure 6.8c is the bar plot of the baseline level of the selected pixel row (highlighted by yellow line in Figure 6.8b). Compared with pixels in PP region, pixels in rSMCF region have a higher baseline level. Pixels near the rSMCF region have mean baseline absorbance in-between.

Apply three clusters Fuzzy-C Mean clustering (maximum number of iterations = 1000, minimum amount of improvement = $10^{-7}$) on data set before any pre-processing. Assigning each cluster with one colour, Figure 6.8b is the pseudo-colour image based on the clustering results. Comparing with the optical image (6.9a), pixels on rSMCF region, pixels on PP region and pixels on near rSMCF region were grouped into different clusters. This result indicated that the rSMCF and is easily distinguishable from the PP/SF matrix.

**Surfactant distribution study**

Based on results in Figure 6.8, we have used HD FTIR imaging to identify the chemical composition of the composite mixtures to understand the distribution of the surfactant. To minimise distortions caused by scattering, baseline correction was achieved using the 'msbackadj' function (WINDOWSIZE 60, STEPSIZE 60) within MATLAB. Variations in sample thickness outside areas containing rSMCF were corrected through normalisation of the PP stretch at 1166 $cm^{-1}$, following a nine-step moving average smoothing. *Quartile one* - (1.5 x *inter quartile range*) and *quartile three* + (1.5 x *inter quartile range*) were used as the boundary to cut off outliers.

In Figure 6.9c, the black pixels represent the mean absorbance (abs > 0.3818) in the range 1800 $cm^{-1}$ to 2000 $cm^{-1}$. The optical cut off value is defined based on the FCM clustering results reported in Figure 6.8.

The pseudo-colour image of the PP matrix, based the integral band area of the surfactant (1111-1136 $cm^{-1}$) is shown Figure 6.9c, where rSMCF pixels have been excluded (black) as the influence of scattering on the surfactant absorption could not be excluded. Using the 20, 40, 60 and 80 percentiles of the values as boundaries, the selected pixels were grouped into five clusters. Five different colours were given to pixels in different clusters. From the image, we can see that SF in the sample is not evenly distributed and has a higher concentration in regions close to the rSMCF.

The standard transmission spectrum of SF and VPP are given in Figure 6.9b. The top half of the figure is the comparison of standard SF spectrum with subtraction of cluster four and one mean spectrum. The mean spectra subtraction matches that of the standard SF spectrum, with peak maxima at 1115 and 1147 $cm^{-1}$. This indicates that the major difference between clusters are SF composition. The green and blue colour spectra in Figure 6.9b are the mean spectrum of each cluster. From cluster one to cluster five, the mean spectra between wavenumber range 1050 and 1170 $cm^{-1}$ increases. This IR spectral region is exactly aligned with SF spectrum peak region (1050-1170 $cm^{-1}$). This indicates the difference between cluster is the SF composition percentage.

A control sample, containing only SF and PP (VPP + 2.5%SF) and imaged using the same method as the VPP + 2%rSMCF + 1%SF sample was obtained to reveal the influence of the rSMCF on the distribution of the SF in the com-

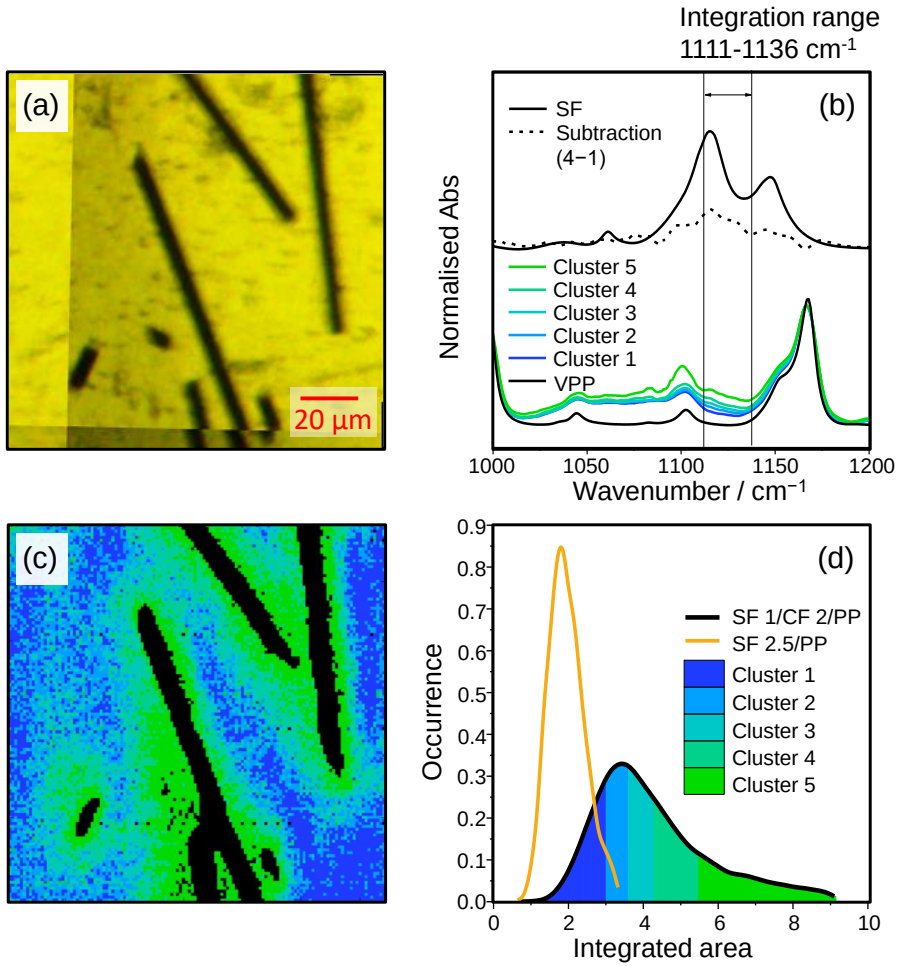



**Figure 6.9:** Images of VPP + 2%rSMCF + 1%SF. (a) Optical image of the selected area for IR spectroscopy examination showing two major compositions: PP and rSMCF. (b) Bottom: FTIR spectrum of PP (black) and the median spectra of the pixel clusters 1-5. Top: FTIR spectrum of SF and the subtraction result of cluster 4 minus cluster 1. The mean spectrum subtraction between cluster 4 and 1 is highly similar to the pure SF spectrum. (c) Pseudo-colour image based on FTIR imaging results. Each block represents a spectrum acquired at this position. Pixels of the same colour indicate the spectra at those positions fall within the same cluster. (d) Integral band area 1111-1136 $cm^{-1}$ absorbance histogram for VPP + 2%rSMCF + 1%SF (excluding regions of rSMCF, black line) and VPP + 2.5%SF (yellow line). The closer the pixel cluster is to rSMCF, the higher the SF composition is.

posite materials. It can be clearly observed from the SF peak integration value density distribution, which is shown in Figure 6.9d. The distribution of the rSMCF sample (VPP + 2%rSMCF + 1%SF, Figure 6.9d, black line) has a mean ($\mu$ = 4.2676) and standard deviation ($\sigma$ = 1.5556). The distribution shows that SF is not evenly distributed due to the influence of the rSMCF. However, the control sample ($SF_{2.5}$/PP, Figure 6.9d, yellow line) has a normal distribution around the mean ($\mu$ = 1.9507) and a lower standard deviation ($\sigma$ = 0.4781), indicating that the SF is more evenly distributed throughout the PP matrix. The lower mean VPP + 2.5%SF compared to VPP + 2%rSMCF + 1%SF is a result of poor mixing between PP and the surfactant in the absence of rSMCF, and so the surfactant is not incorporated into the PP matrix.

### 6.3.2 Virgin polypropylene, maleic anhydride grafted polypropylene with carbon fibre

The polymer reinforcing potential of the recycled fibre can be increased by improving the interfacial adhesion between the fibre and host matrix by the addition of coupling agents. Maleic anhydride grafted polypropylene (MAPP) proved to be an effective compatibiliser in fibre reinforced PP plastics [48, 67]. The long molecular chain of MAPP is responsible for chain entanglements and co-crystallisation with the non-polar PP matrix, which provide mechanical integrity to the host material [68]. Also, the maleic anhydride (MA) groups chemically interact with the functional groups on the fibre surface, resulting in enhanced bonding with the PP matrix. A significant amount of hydroxyl groups on the rCF surface can be observed under X-ray photoelectron spectroscopy

analysis of VPP + MAPP + rCF indicates that the carbon fibre might be covalently bonded with the MAPP through the esterification reaction [69]. As shown in the last section, the distribution of surfactant in the plastic film is highly correlated with the location of rSMCF, in this Section, 20%MAPP is added into the sample, and the distribution of surfactant, as well as the MAPP were studied by FTIR imaging.

The maleic anhydride (MA) grafted rate of the MAPP material is 0.80 wt.%. mn and mw of MAPP are 29,677 and 137,618 respectively.

FTIR imaging in the region of interest was performed on 72.5% VPP, 20% MAPP, 5% recycled short milled carbon fibre (rSMCF) and 2.5% surfactant (VPP + 20%MAPP + 5%rSMCF + 2.5%SF) sample. High spectral resolution, 1 $cm^{-1}$ (wavenumber step size 0.5 $cm^{-1}$) was used. The software returned 64 $\times$ 64 pixels representing the chemical information in the region of interest. Transmission in 'High magnification' mode was used.

Using the method described in Section 6.3.1, the rSMCF was identified by Mie scattering due to the micrometre size rSMCF. Mean absorbance in the wavenumber range 1800-2000 $cm^{-1}$ was used to representing the scattering level of the spectra. The same cut off value as used in the VPP + rSMCF sample analysis 0.3818 was selected. After cutting off the rSMCF, the distortions caused by scattering was corrected by the baseline correction method 'msbackadj' (WINDOWSIZE 60, STEPSIZE 60). Variation in sample thickness outside rSMCF area was corrected through normalisation of the PP stretch at 1166 $cm^{-1}$, following a nine step moving average smoothing.

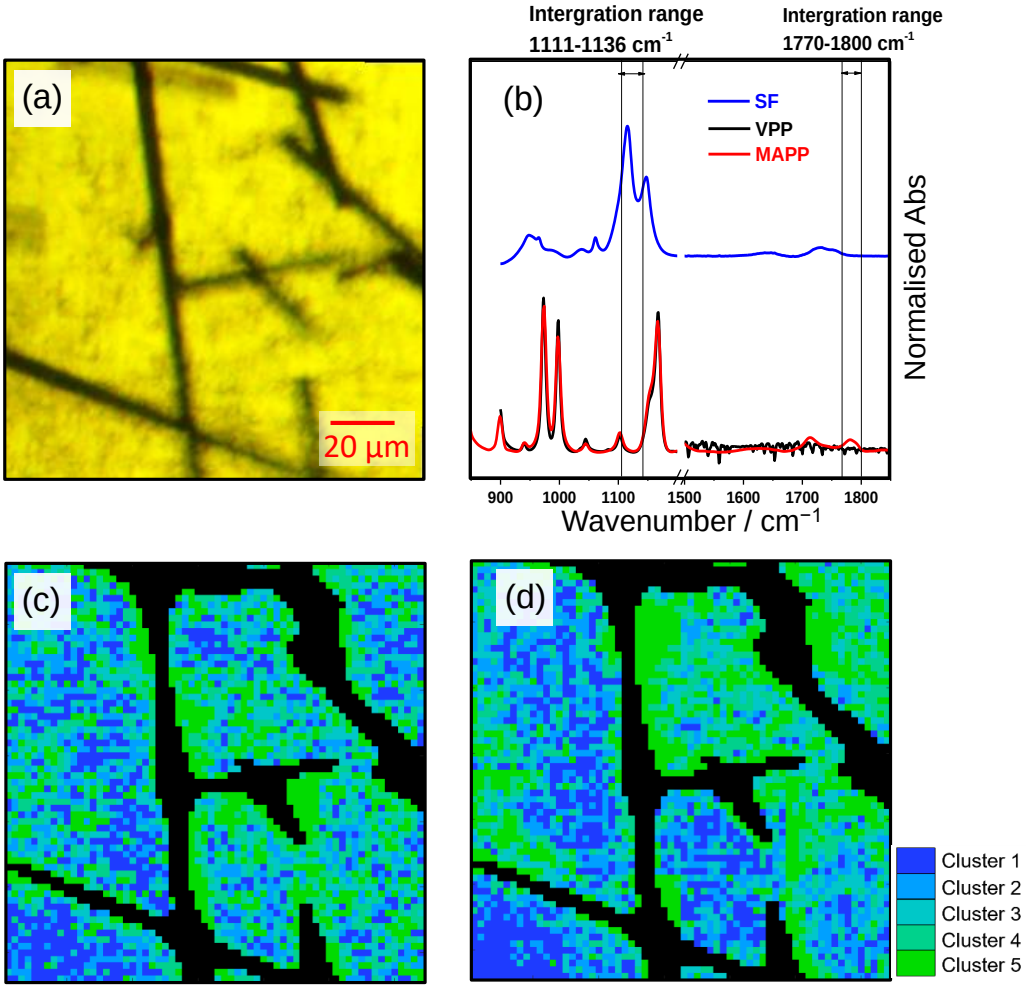



**Figure 6.10:** Images of VPP + 20%MAPP + 5%rSMCF + 2.5%SF. (a) Optical image of the selected area for IR spectroscopy examination. (b) Spectra of VPP, MAPP and SF, with integration bands highlighted. Bottom: FTIR spectrum of PP (black) and MAPP (red). Top: FTIR spectrum of SF (blue). (c) Pseudo-colour image based on MAPP peak integration value, 1770-1800 $cm^{-1}$. Using the 20, 40, 60 and 80 percentiles of the integration values as boundaries, pixels were grouped into five clusters. Five different colours were given to pixels in different clusters. (d) Pseudo-colour image based on SF peak integration, 1111-1136 $cm^{-1}$. The same clustering and colourmap as (c) were applied. The distribution of MAPP shows less correlation with the rSMCF location.

MAPP has similar IR spectrum with VPP, except additional peaks around 3500 $cm^{-1}$ and multi-peaks at 1500-1800 $cm^{-1}$ range, due to the O-H and C=O vibration of the MA group [47]. Figure 6.10a is the optical image of the IR spectroscopy examined area. Figure 6.10b is the IR spectra of VPP, MAPP and SF. Considering the SF peaks in range 1600 - 1770 $cm^{-1}$, range 1770-1800 $cm^{-1}$ was used as the integration range for MAPP identification. The same SF integration range as VPP + rSMCF sample analysis in Section 6.3.1 was used to identify the SF composition. Integration ranges used to produce the pseudo-colour images are highlighted in Figure 6.10b.

Figure 6.10d is the pseudo-colour image based on the integral band area of the SF (1111-1136 $cm^{-1}$), where rSMCF pixels have been excluded (black). Using the 20, 40, 60 and 80 percentiles of the integral values as boundaries, the selected pixels were grouped into five clusters. Five different colours were given to pixels in different clusters. The higher the cluster number, the higher the surfactant peak integration value. A highly uneven distribution of SF can be observed in the resulting image. Pixels nearby to the rSMCF have higher SF composition. A clear cluster of high SF composite can be observed in the top right conner of the image, where a cluster of rSMCF located. These results agree well with the VPP + rSMCF + SF analysis result (Figure 6.9, on page 135). The distribution of SF is highly relevant with the rSMCF location.

Figure 6.10c is the pseudo-colour image based on the integral band area of the MAPP (1770-1800 $cm^{-1}$). Unlike the distribution of SF, the distribution of MAPP has less correlation with the location of rSMCF. The region near the rSMCF, as well as some part of the PP matrix contains high-level of MAPP.

The significant amount of MA groups in the region near the rSMCF might be the result of chemical bonding between fiber and MAPP [69]. The MAPP located in the middle of PP matrix is responsible for chain entanglements and co-crystallisation with the non-polar PP matrix [68].

Compared with MAPP, SF is highly unevenly distributed in the resulting film. This confirms the hypothesis that, for rSMCF reinforced PP plastic, SF attaching on or locating nearby the rSMCF may be one of the major reasons for low interfacial adhesion between rSMCF and PP matrix.

### 6.3.3 Acrylonitrile butadiene styrene with Calcium carbonate

Employing inorganic particulate fillers, such as calcium carbonate ($CaCO_3$) [70], kaolin and glass beads [71], and talcum powder [72], to improve properties and/or lower costs of polymer products has been a common practise [73, 74]. Acrylonitrile butadiene styrene (ABS) is an engineering resin used extensively in industry owing to its good mechanical, optical and processing properties. $CaCO_3$ is commonly used to lower the cost of ABS [73]. Mechanical property study of ABS/CaCO3 composites [75] revealed that by adding $CaCO_3$, the composite shows an increase in tensile modulus compared with neat ABS. Agglomeration is a common problem in thermoplastics filled with $CaCO_3$ particles [75], due to the inadequate adhesion between $CaCO_3$ particles and resins. It is well known that shape and size of the agglomeration strongly affect the properties of the final product [75, 76]. In this investigation, FTIR imaging was proposed as an imaging tool to visualise the agglomeration problem. The FTIR imaging data of an ABS + $CaCO_3$ film was collected. The distribution of $CaCO_3$ particles

in the ABS matrix was visualised in both real-value and pseudo-colour based imaging techniques.

FTIR imaging was performed on 80% ABS, 20% $CaCO_3$ (ABS + 20%$CaCO_3$) sample using the following setting: pixel size 1.1 $\mu m^2$, spectral resolution 4 $cm^{-1}$ and co-adding 16 scans. Using the film making procedure developed, the thickness of the sample varies from 15 to 20 $\mu$m across the sample (the lowest and the highest value of those five measurements).

'msbackadj' function (WINDOWSIZE 60, STEPSIZE 60) baseline correction and nine-point moving average smooth algorithm, which is same with pre-processing methods used in VPP + talc sample in Section 6.2, were applied.

Figure 6.11 are the resulting image of ABS + 20%$CaCO_3$ sample. The standard spectra of ABS and $CaCO_3$ are given in Figure 6.11b with overlapping peak region excluded. Peak at 871 $cm^{-1}$ was selected as the $CaCO_3$ peak, while peak at 3026 $cm^{-1}$ was selected to represent the ABS composition. With the absorbance value of the peak directly proportional to the composition concentration, the peak ratio A(871 $cm^{-1}$)/A(3026 $cm^{-1}$) was used to represent the $CaCO_3$/ABS ratio in the composite. Each block representing a spectrum acquired at this position, a real-value image (Figure 6.11c) based on peak ratio was produced, where outliers (pixels in the high and low range) were cut off, using the standard outlier identification method introduced in Section 6.3.1. A highly uneven distribution of $CaCO_3$ can be observed in Figure 6.11c. The bottom right corner of the image has the most yellow/red (high peak ratio level) pixels. The top-half of the image is dominated by blue (low peak ratio level) colour pixels.

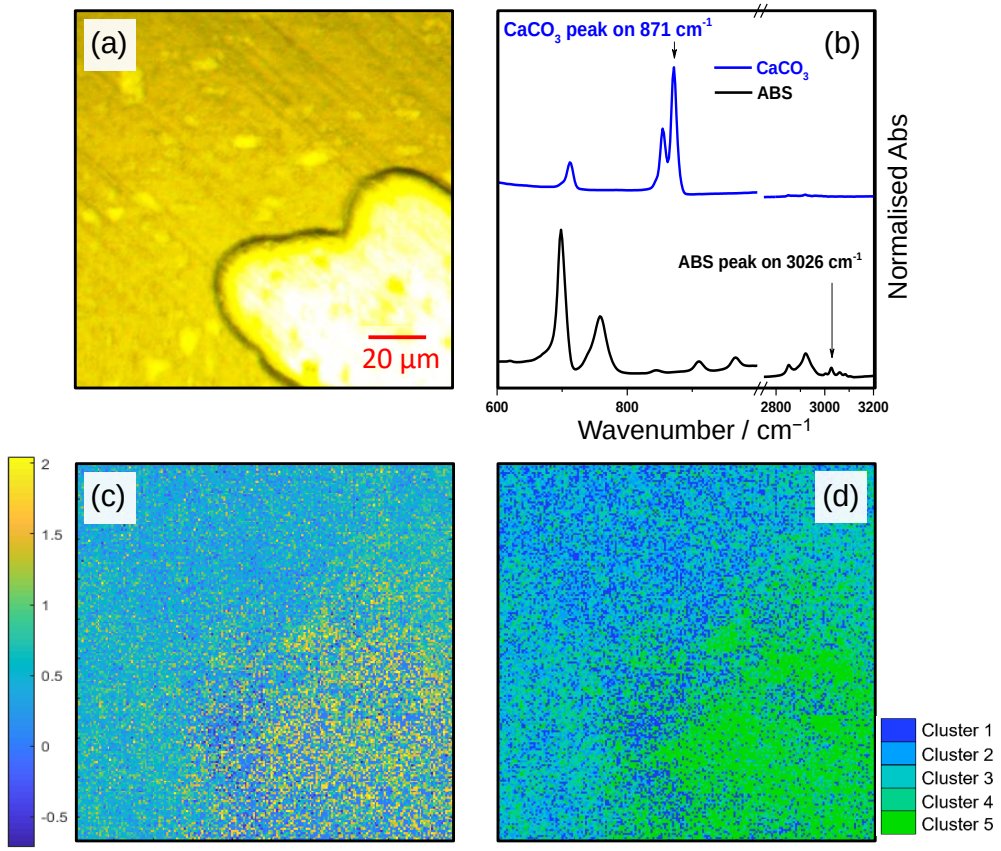



**Figure 6.11:** Images of ABS + 20%$CaCO_3$ sample. (a) Optical image of the selected area for IR spectroscopy examination. (b) Spectra of ABS (black line) and $CaCO_3$ (blue line), with representative peak highlighted. (c) Real-value image based on peak ratio A(871 $cm^{-1}$)/A(3026 $cm^{-1}$). Each block represents a spectrum acquired at this position. Different colour were given to each block based on the peak ratio value. The colour map used is given on the left of the image. (d) Pseudo-colour image based on peak ratio clustering results. Pixels of the same colour indicate the spectra at those positions fall within the same cluster. 20, 40, 60 and 80 percentiles of the peak ratio value were used as boundaries between clusters. Agglomeration of $CaCO_3$ in the bottom right corner of the sample can be clearly observed.

To visualise the agglomeration, clustering method was applied to group spectra into clusters. Using the 20, 40, 60 and 80 percentiles of the values as boundaries, pixels were grouped into five clusters. Five different colours were given to pixels in different clusters. From Figure 6.11d, we can see that the bottom right corner of the sample contains pixels from the green cluster (high $CaCO_3$/PP ratio cluster), while the top left corner has mostly pixles belonging to blue cluster (low $CaCO_3$/PP ratio cluster). The agglomeration of $CaCO_3$ in the bottom right corner of the sample can be clearly observed..

Comparing with the optical image (Figure 6.11a), the high $CaCO_3$ region (bottom right corner in Figure 6.11d) corresponds well with the 'bright region' of the optical image. The increasing of $CaCO_3$ filler composition in bottom right corner caused the transmission optical properties change, which is causing the 'bright region' under optical microscope.

Applying two-cluster Fuzzy C-Mean Clustering on the whole spectrum of the raw data set, the 'bright region' in Figure 6.11a was successfully identified and separated into a different cluster with the rest of the sample region. Comparison of spectra between clusters shows that the "bright region' cluster spectra have higher baseline. The region contains more $CaCO_3$ has high baseline compared with low $CaCO_3$ composition region. The reason for this baseline problem can be either over-thickness (caused by physical properties change due to higher $CaCO_3$ composition) or higher scattering caused by the $CaCO_3$ particles in that region. The Fuzzy C-Mean Clustering method used and the phenomenon observed is exactly the same with VPP + 20%talc sample reported in Figure 6.7 on page 129.

### 6.3.4 Virgin polypropylene with poly(ethylene terephthalate)

Combining the excellent properties of more than one polymer, polymer blending is a convenient route for the development of new materials [77]. During the blending process a large variety of morphologies of the dispersed phase can be formed, e.g. spheres or ellipsoids, fibrils or plates [78–80], depending on the weight ratio of the blend components, their chemical structure, their properties, and the processing conditions. Blend properties are strongly dependent on the state of mixing. Optical microscopy [81] or scanning electron microscopy (SEM) [81, 82] of blends are typically used in dispersed phase studies. Poly(ethylene terephthalate) (PET), as one of the commercially important polymers, has been most extensively studied [14, 83]. The combination of PP with PET offers some advantages over the pure components [81, 84–86]. PET may enhance the stiffness of PP at higher temperatures while the polyolefin could facilitate crystallisation of PET by heterogeneous nucleation further raising blend stiffness. The lower permeability of PET towards water vapour and oxygen could be usefully utilised in packaging materials. In this study, FTIR imaging was used to exam the PP/PET blend to visualise the disperse phase of PET. The PET material used in this study is produced by Dupont, PT 2251.

FTIR imaging was performed on 80% PP, 20% PET (PP + 20%PET) sample using the following settings: pixel size 1.1 $\mu m^2$, spectral resolution 4 $cm^{-1}$ and co-adding 16 scans. The thickness of the sample varies from 18 to 20 $\mu$m across the sample (the lowest and the highest value of those five measurements). A wider field of view was also investigated using the same setting except the pixel

size was changed to 5.5 $\mu$m$^2$. 'msbackadj' function (WINDOWSIZE 60, STEP-SIZE 60) baseline correction and nine-point moving average smooth algorithm, which are same with pre-processing methods used in Section 6.2, were applied.

Figure 6.12 are the resulting images of PP + 20%PET sample on the high spatial resolution. The standard spectra of PP and PET are given in Figure 6.12b with overlapping peak region excluded. Peak at 1713 cm$^{-1}$ was selected as the PET peak [87], while peak at 2835 cm$^{-1}$ was selected to represent the PP composition. With the absorbance value of the peak directly proportional to the composition concentration, the peak ratio A(1713 cm$^{-1}$)/A(2835 cm$^{-1}$) was used to represent the PET/PP ratio in the composite. Each block representing a spectrum acquired at this position, a real-value image (Figure 6.12c) based on peak ratio was produced, where outliers (pixels in the high and low range) were identified and given the boundary values. The real-value image (Figure 6.12c) corresponds well with the optical image in Figure 6.12a. PET disperse phase can be clearly observed. PET was highly uneven distributed among the PP matrix.

Clustering method was applied to group spectra into five clusters. The pseudo-colour image of the PET matrix is shown in Figure 6.12d. PET agglomeration can be clearly observed. Large areas of the sample are either compose extremely high (cluster 5) or extremely low PET (cluster 1) concentration. The melting point for PET is 250-260 $^o$C, while PP has a melting point of 130-171 $^o$C. The PET disperse phase indicates that the uneven distribution may due to the PET was not fully melted before the two materials were mixed.

The same pattern can be observed on a larger scale image using 5.5 $\mu$m$^2$ pixel

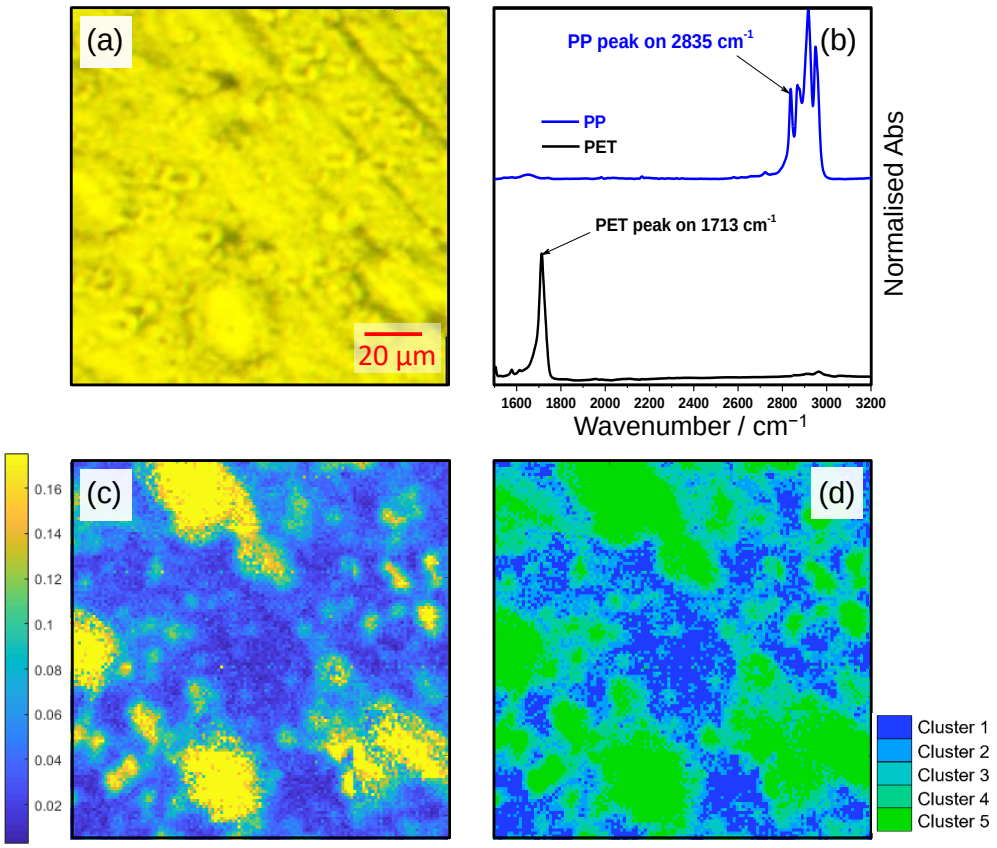



**Figure 6.12:** Images of PP + 20%PET sample under 5.5 $\mu m^2$ spatial resolution. (a) Optical image of the selected area for IR spectroscopy examination. (b) Spectra of PET (black line) and PP (blue line), with representative peak highlighted. (c) Real-value image based on peak ratio A(1713 $cm^{-1}$)/A(2835 $cm^{-1}$). Each block represents a spectrum acquired at this position. Different colour were given to each block based on the peak ratio value. The colour map used is given on the left of the image. (d) Pseudo-colour image based on peak ratio clustering results. Pixels of the same colour indicate the spectra at those positions fall within the same cluster. 20, 40, 60 and 80 percentiles of the peak ratio value were used as boundaries between clusters.

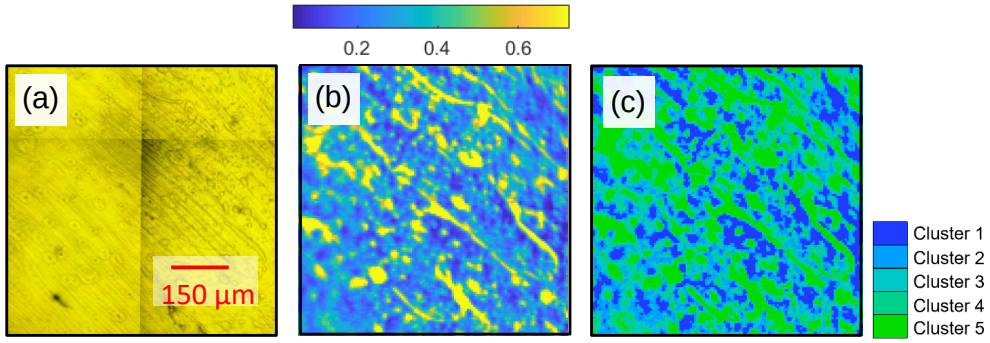



**Figure 6.13:** Images of VPP + 20%PET under 5.5 $\mu m^2$ spatial resolution. (a) Optical image of the selected area for IR spectroscopy examination. (b) Real-value image based on peak ratio A(1713 $cm^{-1}$)/A(2835 $cm^{-1}$). The colour map used is given on the top of the image. (c) Pseudo-colour image based on peak ratio clustering results. 20, 40, 60 and 80 percentiles of the peak ratio value were used as boundaries between clusters.

size (Figure 6.13). Compared with the spheres or plates disperse shape observed in Figure 6.12, ellipsoids and fibrils as well as spheres and plates shapes can be seen in the larger scale imaging. Either spatial resolution choice has advantages and disadvantages. With large region of the sample imaged, low spatial FTIR imaging has higher representativeness. However, it is limited in providing enough details of the distribution. High spatial resolution FTIR imaging can provide real micron scale imaging of the blend based on individual components, but the selecting of the interesting region to study is subject to the choice of individual researcher. The analysis results have less representativeness of the whole sample.

The highly uneven distribution of PET observed indicated the low performance of blending, which could lead to the dropping performance of the resulting plastic products The phenomenon implies the PET was not fully melted before

the mixing. Longer melting time or higher melting temperature is suggested. This study reveals the great potential of FTIR imaging on providing insight into the blending studies.

### 6.3.5 Section conclusion

Utilising FTIR imaging, we have examined four of the most widely used polymer composites. The analysis of the rSMCF reinforced PP indicated the surfactant used in the rSMCF preparation remains in the resulting composite, which is causing a weak inter-facial interaction between fibres and PP matrix. Comparing the surfactant distribution with the coupling agent MAPP distribution, we found the latter is less correlated with the position of fibres and relatively more even distributed in the PP matrix. This agrees with the fact that the surfactant bonds with the fibre during the dispersal which can be hard to break. Uneven distribution were observed in both the analysis result of $CaCO_3$ reinforced ABS and PP/PET copolymer which indicate the poor performance of the mixer. In the $CaCO_3$ case, a better blending method is required; while in the copolymer case, a better melting method is needed. With the uneven distribution of composition observed, the causing of the uneven distribution obtained, imaging technique in FTIR spectroscopy not only provide micron level spatial information of the composition but also solutions of improving the inter-facial interactions between compositions. Those pilot studies provide insights into applying FTIR imaging for plastic sample analysis. FTIR imaging has promising potential in plastic sample analysis, thus improve the physical/chemical performance of the plastic products.

CHAPTER 7

# Conclusion and future work

This study provides a valuable insight into the application of FTIR spectroscopy for identification and characterisation of the individual components present in materals. The calibration of talc concentration with IR spectrum peak ratio/integration and the application of the calibration result on a film sample analysis demonstrated FTIR spectroscopy as a powerful tool for recycled plastic analysis. Using FTIR imaging, we have examined the chemical composition of: ABS + $CaCO_3$, PP + talc, PP + rSMCF + SF, PP + MAPP + rSMCF + SF and PP + PET on the micron scale. The results indicate that, with data analysis, the chemical composition distribution in the polymer films can be revealed. Robust spectral and spatial information can be obtained using FTIR imaging.

In general, this study confirms the great potential of using FTIR spectroscopy as an analytical tool for process control during the plastic recycling and to use high-definition FTIR imaging to monitor the chemical composition and spatial distribution of plastic composites. To achieve the ultimate goal of online analysis and sorting of plastic using FTIR spectroscopy, with this pilot study, we

believe further works in this project should be focusing on the following directions.

### 7.0.1 Spectra library

Even though, abundant literature can be found in terms of IR spectroscopy study of plastic samples. A systematic spectra library is still needed to be able to analysis recycled plastics. For domestic wastes, mainly five types of polymers are used [6]. The types of filler/additives are quite limited. It is a practical method to obtain a spectra library of most commonly available domestic plastic, and sorting the plastic based on the IR spectrum obtained from the waste.

Technical products, on the other hand, use a much wider range of plastic and filler/additives to gain different physical/chemical features of plastic. Many of the waste plastics are copolymers or blends. It is difficult to obtain a spectra library covering majority of the different combination or ratio of compositions. It is, however, possible to use FTIR spectroscopy on the evaluation of one type of industrial products. Automobile part production, for example, normally use PP using talc as filler. Different manufacturers use different amounts of talc, which is causing the low quality of the recycled products of that plastic. Evaluation of the talc percentage using FTIR spectroscopy can be beneficial. A preliminary study of building a spectra library with standard spectrum of PP + talc in the different ratio and further analysis of un-known composition material is given in Section 6.2. Further application this method to other industrial plastics are needed.

### 7.0.2 Degradation and oxidation

When overheated, the energy input from the heat can cause the lose of a hydrogen atom from the polymer chain. This creates a 'free radical' (R·) polymer and a hydrogen atom with an unpaired electron (H·). The free radical (R·) can react with oxygen in the atmosphere to form a peroxy radical (ROO·), which can then remove a hydrogen atom from another polymer chain to form a hydroperoxide (ROOH) and so regenerate a free radical (R·) again. The newly generated free radical (R·) can repeat the whole process again. This process can be accelerated causing molecular deterioration of the polymer, which is termed as polymer thermal ageing and degradation [88].

Thermal ageing and degradation can cause the polymer properties to deteriorate significantly [89]. Online thermal ageing/photo-oxidisation detection using FTIR spectroscopy can improve the reliability of recycled plastic for high-end applications.

### 7.0.3 Industrial transformation

The developing of a continuous plastic sorting system using FTIR spectroscopy as the online analysis tool is the ultimate goal. With the supporting of those fundamental works stated above, online detection and sorting of plastic sample for accurate recycling can be adopted in the manufacturing sector. Technology transfer from the lab to the plastic recycling industry is needed before the goal can be achieved.

# Bibliography


[1] *Polymer identification using mid-infrared spectroscopy*. Tech. rep. PerkinElmer, 2011.

[2] Charles Moore. *Plastic pollution*. 2018. URL: https://www.britannica.com/science/plastic-pollution (visited on 02/11/2018).

[3] J. C. Arnold et al. "The use of FTIR mapping to assess phase distribution in mixed and recycled WEEE plastics". In: *Polym. Test* 29.4 (2010), pp. 459–470.

[4] Todd Johnson. *Recycling Plastics*. Ed. by ThoughtCp. 9th Apr. 2017. URL: https://www.thoughtco.com/recycling-plastics-820356 (visited on 17/12/2018).

[5] Philip Dewey. *This is what happens to your plastic after you put it in the recycling, and why it's going to Turkey*. Ed. by Wales Online. 7th Oct. 2017. URL: https://www.walesonline.co.uk/news/wales-news/what-happens-your-plastic-after-13728985 (visited on 27/05/2019).

[6] Norbert Eisenreich and Thomas Rohe. *Infrared Spectroscopy in Analysis of Plastics Recycling*. Wiley Online Library, 2000.

[7] George Socrates. *Infrared and Raman characteristic group frequencies: tables and charts*. John Wiley & Sons, 2004.

[8] Daimay Lin-Vien et al. *The handbook of infrared and Raman characteristic frequencies of organic molecules*. Elsevier, 1991.

[9] John M. Chalmers and Neil J. Everall. "Qualitative and quantitative analysis of plastics, polymers and rubbers by vibrational spectroscopy". In: *Handbook of vibrational spectroscopy*. 2006.

[10] Jack L. Koenig, L. d'Esposito and M. K. Antoon. "The ratio method for analyzing infrared spectra of mixtures". In: *Appl. Spectrosc.* 31.4 (1977), pp. 292–295.

[11] Hugh E. Diem and Samuel Krimm. "Quantitative Analysis of Binary Mixtures. I. General Method". In: *Appl. Spectrosc.* 35.4 (1981), pp. 421–428.

[12] S. C. Y. L. Krimm, C. Y. Liang and G. B. B. M. Sutherland. "Infrared spectra of high polymers. II. Polyethylene". In: *J. Chem. Phys.* 25.3 (1956), pp. 549–562.

[13] Heinz Wilhelm Siesler et al. "Infrared and Raman spectroscopy of polymers". In: *Practical spectroscopy;* 1980.

[14] I. M. Ward and M. A. Wilding. "Infra-red and Raman spectra of poly (m-methylene terephthalate) polymers". In: *Polymer* 18.4 (1977), pp. 327–335.

[15] L. R. Schroeder and Stuart L. Cooper. "Hydrogen bonding in polyamides". In: *J. Appl. Phys.* 47.10 (1976), pp. 4310–4317.

[16] Jack L. Koenig. "Fourier transform infrared spectroscopy of polymers". In: *Spectroscopy: NMR, Fluorescence, FT-IR*. Springer, 1984, pp. 87–154.

[17] Kenneth C. Cole. "Vibrational spectroscopy of polymer composites". In: *Handbook of Vibrational Spectroscopy*. 2006.

[18] S. Mordechai et al. "Early diagnosis of Alzheimer's disease using infrared spectroscopy of isolated blood samples followed by multivariate analyses". In: *Analyst* (2017).

[19] Michael M. Coleman et al. "Hydrogen bonding in polymers. 4. Infrared temperature studies of a simple polyurethane". In: *Macromolecules* 19.8 (1986), pp. 2149–2157.

[20] Dorab E. Bhagwagar et al. "Mapping the phase diagram of an aromatic polyamide blend using infrared spectroscopy". In: *J. Polym. Sci., Part B: Polym. Phys.* 29.12 (1991), pp. 1547–1558.

[21] Michael M. Coleman, George J. Pehlert and Paul C. Painter. "Functional group accessibility in hydrogen bonded polymer blends". In: *Macromolecules* 29.21 (1996), pp. 6820–6831.

[22] George J. Pehlert et al. "Self-association versus interassociation in hydrogen bonded polymer blends: 2. Comparison of theoretical and experimental miscibility windows for poly (2, 6-dialkyl-4-vinyl phenol) blends". In: *Polymer* 37.21 (1996), pp. 4763–4771.

[23] P. C. Painter et al. "Intramolecular screening effects in polymer mixtures. 1. Hydrogen-bonded polymer blends". In: *Macromolecules* 30.4 (1997), pp. 932–942.

[24] Andrew Garton. *Infrared spectroscopy of polymer blends, composites and surfaces*. Hanser Munich etc., 1992.

[25] Takeshi Ohwaki and Hatsuo Ishida. "Infrared Spectroscopy for Composites." In: *Characterization of Composite Materials*. Elsevier, 1994, pp. 184–271.

[26] E. T. Thostenson et al. "Carbon nanotube/carbon fiber hybrid multiscale composites". In: *J. Appl. Phys.* 91.9 (2002), pp. 6034–6037.

[27] Enrique J. Garcia et al. "Fabrication and multifunctional properties of a hybrid laminate with aligned carbon nanotubes grown in situ". In: *Compos. Sci. Technol.* 68.9 (2008), pp. 2034–2041.

[28] C. Sellitti, J. L. Koenig and H. Ishida. "Surface characterization of carbon fibers and interphase phenomena in epoxy-reinforced composites". In: *Mater. Sci. Eng. A* 126.1-2 (1990), pp. 235–244.

[29] C. Sellitti, J. L. Koenig and H. Ishida. "Surface characterization of graphitized carbon fibers by attenuated total reflection Fourier transform infrared spectroscopy". In: *Carbon* 28.1 (1990), pp. 221–228.

[30] C. Sellitti, J. L. Koenig and H. Ishida. "Enhancement of the Absorptions of Surface Species Present on Graphitized Carbon Fibers: A Theoretical Approach". In: *Appl Spectrosc* 44.5 (1990), pp. 830–836.

[31] C. Sellitti, J. L. Koenig and H. Ishida. "Influence of the surface treatment of graphitized carbon fibers on the curing of an amine-catalyzed epoxy-anhydride system". In: *J. Polym. Sci. Pol. Phys.* 28.7 (1990), pp. 1121–1135.

[32] Carol Kozlowski and Peter M. A. Sherwood. "X-ray photoelectron spectroscopic studies of carbon-fibre surfaces. Part 4.—The effect of electrochemical treatment in nitric acid". In: *J. Chem. Soc. Faraday Trans.* 80.8 (1984), pp. 2099–2107.

[33] R. L. Zhang et al. "Directly grafting graphene oxide onto carbon fiber and the effect on the mechanical properties of carbon fiber composites". In: *Mater. Des.* 93 (2016), pp. 364–369.

[34] B. Van Mele and E. Verdonck. "Physico chemical characterization of the fibre matrix interaction in polyethylene fibre epoxy matrix composites". In: *Compos. Interfaces* 3.2 (1995), pp. 101–119.

[35] Lin Tu and Deon Kruger. "Engineering properties of epoxy resins used as concrete adhesives". In: *Materials* 93.1 (1996), pp. 26–35.

[36] R. Anbarasan et al. "Chemical grafting of aniline and o-toluidine onto poly (ethylene terephthalate) fiber". In: *J. Appl. Polym. Sci.* 73.1 (1999), pp. 121–128.

[37] Miguel Arroyo, Mara Tejera and Luis Ibarra. "Functionalization of polyester PET fibers". In: *J. Appl. Polym. Sci.* 48.6 (1993), pp. 1009–1014.

[38] O. Durcova, P. Michlik and L' Knotek. "Evaluation of the calcium carbonate dispersion in polypropylene composite fibres". In: *Polym. Test* 8.4 (1988), pp. 269–275.

[39] Dieter O. Hummel and Friedrich K. Scholl. *Infrared analysis of polymers, resins and additives: an atlas [by] Hummel-Scholl*. Vol. 1. Wiley-Interscience, 1971.

[40] John M. Chalmers. "Infrared spectroscopy in analysis of polymers and rubbers". In: *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*. 2006.

[41] Prabuddha Mukherjee et al. "Relating Post-yield Mechanical Behavior in Polyethylenes to Spatially Varying Molecular Deformation Using Infrared Spectroscopic Imaging: Homopolymers". In: *Macromolecules* (2018).

[42] Rohit Bhargava and Ira W Levin. *Spectrochemical analysis using infrared multichannel detectors*. John Wiley & Sons, 2008.

[43] A. Jayakrishnan and M. C. Sunny. "Phase transfer catalysed surface modification of plasticized poly (vinyl chloride) in aqueous media to retard plasticizer migration". In: *Polymer* 37.23 (1996), pp. 5213–5218.

[44] L. B. Brennan, D. H. Isaac and J. C. Arnold. "Recycling of acrylonitrile–butadiene–styrene and high-impact polystyrene from waste computer equipment". In: *J. Appl. Polym. Sci.* 86.3 (2002), pp. 572–578.

[45] Xiaojuan Bai, D. H. Isaac and Keith Smith. "Reprocessing acrylonitrile–butadiene–styrene plastics: structure–property relationships". In: *Polym. Eng. Sci.* 47.2 (2007), pp. 120–130.

[46] Harald Martens and Edward Stark. "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy". In: *J. Pharm. Biomed. Anal.* 9.8 (1991), pp. 625–635.

[47] Yuan Liu et al. "Bio-inspired polydopamine functionalization of carbon fiber for improving the interfacial adhesion of polypropylene composites". In: *RSC Adv.* 5.130 (2015), pp. 107652–107661.

[48] K. H. Wong et al. "Effect of coupling agents on reinforcing potential of recycled carbon fibre for polypropylene composite". In: *Compos. Sci. Technol.* 72.7 (2012), pp. 835–844.

[49] Fateme Rezaei et al. "Development of short-carbon-fiber-reinforced polypropylene composite for car bonnet". In: *Polym.-Plast. Technol. Eng.* 47.4 (2008), pp. 351–357.

[50] Abraham Savitzky and Marcel J. E. Golay. "Smoothing and differentiation of data by simplified least squares procedures." In: *Anal. Chem.* 36.8 (1964), pp. 1627–1639.

[51] Matthew J. Baker et al. "Using Fourier transform IR spectroscopy to analyze biological materials". In: *Nat. Protoc.* 9.8 (2014), pp. 1771–1791.

[52] Júlio Trevisan et al. "Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives". In: *Analyst* 137.14 (2012), pp. 3202–3215.

[53] Benjamin Bird and Jeremy Rowlette. "High Definition Infrared Chemical Imaging of Colorectal Tissue using a Spero QCL Microscope". In: *Analyst* (2017).

[54] Hyunjin Shin and Mia K Markey. "A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples". In: *J. Biomed. Inform.* 39.2 (2006), pp. 227–248.

[55] Francis L. Martin et al. "Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample". In: *J. Comput. Biol.* 14.9 (2007), pp. 1176–1184.

[56] Rohit Bhargava and Anant Madabhushi. "Emerging themes in image informatics and molecular analysis for digital pathology". In: *Annu. Rev. Biomed. Eng.* 18 (2016), pp. 387–412.

[57] Witold K. Surewicz, Henry H. Mantsch and Dennis Chapman. "Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment". In: *Biochemistry* 32.2 (1993), pp. 389–394.

[58] José Britti Bacalhau, Túlio Mumic Cunha and Conrado Ramos Moreira Afonso. "Effect of Ni content on the Hardenability of a Bainitic Steel for Plastics Processing". In: *COBEM-2017*. 2017.

[59] Géraldine Oliveux, Luke O. Dandy and Gary A. Leeke. "Current status of recycling of fibre reinforced polymers: Review of technologies, reuse and resulting properties". In: *Prog. Mater Sci.* 72 (2015), pp. 61–99.

[60] A. K. Bledzki, S. Reihmane and J. Gassan. "Thermoplastics reinforced with wood fillers: a literature review". In: *Polym -Plast Technol. Eng.* 37.4 (1998), pp. 451–468.

[61] M. Khalid et al. "Comparative study of polypropylene composites reinforced with oil palm empty fruit bunch fiber and oil palm derived cellulose". In: *Mater. Des.* 29.1 (2008), pp. 173–178.

[62] Michael Q. Tran et al. "Carbon fibre reinforced poly (vinylidene fluoride): impact of matrix modification on fibre/polymer adhesion". In: *Compos. Sci. Technol.* 68.7-8 (2008), pp. 1766–1776.

[63] Fateme Rezaei, R. Yunus and N. A. Ibrahim. "Effect of fiber length on thermomechanical properties of short carbon fiber reinforced polypropylene composites". In: *Mater. Des.* 30.2 (2009), pp. 260–263.

[64] Guozhan Jiang, Stephen J. Pickering and N. A. Warrior. "On the dispersion of recycled carbon fibre for manufacturing random mat using wet-laid process". In: *Proceeding of 16th International Conference on Composite Structures, Porto, Portugal*. 2011.

[65] Walker Camacho and Sigbritt Karlsson. "NIR, DSC, and FTIR as quantitative methods for compositional analysis of blends of polymers obtained from recycled mixed plastic waste". In: *Polym. Eng. Sci.* 41.9 (2001), pp. 1626–1635.

[66] Tatiana Konevskikh et al. "Mie scatter corrections in single cell infrared microspectroscopy". In: *Faraday Discuss* 187 (2016), pp. 235–257.

[67] Nevin Gamze Karsli et al. "Effects of irradiated polypropylene compatibilizer on the properties of short carbon fiber reinforced polypropylene composites". In: *Radiat. Phys. Chem.* 84 (2013), pp. 74–78.

[68] Anand R. Sanadi et al. "Renewable agricultural fibers as reinforcing fillers in plastics: Mechanical properties of kenaf fiber-polypropylene composites". In: *Ind. Eng. Chem. Res.* 34.5 (1995), pp. 1889–1896.

[69] K. H. Wong et al. "Characterisation of recycled carbon fibre: mechanical properties and surface chemistry". In: *SAMPE fall technical conference proceedings: Global advances in materials and process engineering, Dallas, TX: society for the advancement of material and process engineering*. 2006.

[70] Ji-Zhao Liang et al. "Crystalline properties of poly (L-lactic acid) composites filled with nanometer calcium carbonate". In: *Compos. Part. B-Eng.* 45.1 (2013), pp. 1646–1650.

[71] R. E. Lavengood, L. Nicolais and M. Narkis. "A deformational mechanism in particulate-filled glassy polymers". In: *J. Appl. Polym. Sci.* 17.4 (1973), pp. 1173–1185.

[72] D. M. Bigg. "Mechanical properties of particulate filled polymers". In: *Polym. Compos.* 8.2 (1987), pp. 115–122.

[73] Ji-Zhao Liang. "Melt rheology of nanometre-calcium-carbonate-filled acrylonitrile–butadiene–styrene (ABS) copolymer composites during capillary extrusion". In: *Polym. Int.* 51.12 (2002), pp. 1473–1478.

[74] L. Jiang et al. "Strengthening acrylonitrile-butadiene-styrene (ABS) with nano-sized and micron-sized calcium carbonate". In: *Polymer* 46.1 (2005), pp. 243–252.

[75] Aimin Zhang et al. "Study on mechanical and flow properties of acrylonitrile-butadiene-styrene/poly (methyl methacrylate)/nano-calcium carbonate composites". In: *Polym. Compos.* 31.9 (2010), pp. 1593–1602.

[76] P. Cheremisinoff. *Handbook of engineering polymeric materials*. CRC Press, 1997.

[77] C. I. W. Calcagno et al. "The role of the MMT on the morphology and mechanical properties of the PP/PET blends". In: *Compos. Sci. Technol.* 68.10-11 (2008), pp. 2193–2200.

[78] A. Leclair and B. D. Favis. "The role of interfacial contact in immiscible binary polymer blends and its influence on mechanical properties". In: *Polymer* 37.21 (1996), pp. 4723–4728.

[79] D. R. Paul. *Polymer blends*. Vol. 1. Elsevier, 2012.

[80] J. Li and B. D. Favis. "Characterizing co-continuous high density polyethylene/polystyrene blends". In: *Polymer* 42.11 (2001), pp. 5047–5053.

[81] C. P. Papadopoulou and N. K. Kalfoglou. "Comparison of compatibilizer effectiveness for PET/PP blends: their mechanical, thermal and morphology characterization". In: *Polymer* 41.7 (2000), pp. 2543–2555.

[82] K Friedrich et al. "Microfibrillar reinforced composites from PET/PP blends: processing, morphology and mechanical properties". In: *Compos. Sci. Technol.* 65.1 (2005), pp. 107–116.

[83] D. A. Jarvis et al. "Characterization of biaxial orientation in poly (ethylene terephthalate) by means of refractive index measurements and Raman and infra-red spectroscopies". In: *Polymer* 21.1 (1980), pp. 41–54.

[84] P. Bataille, S. Boisse and H. P. Schreiber. "Mechanical properties and permeability of polypropylene and poly (ethylene terephthalate) mixtures". In: *Polym. Eng. Sci.* 27.9 (1987), pp. 622–626.

[85] M. Xanthos, M. W. Young and J. A. Biesenberger. "Polypropylene/polyethylene terephthalate blends compatibilized through functionalization". In: *Polym. Eng. Sci.* 30.6 (1990), pp. 355–365.

[86] Yi-Jun Sun et al. "In situ compatibilization of polypropylene and poly (butylene terephthalate) polymer blends by one-step reactive extrusion". In: *Polymer* 37.18 (1996), pp. 4119–4127.

[87] Ana Paula dos Santos Pereira et al. "Processing and Characterization of PET Composites Reinforced With Geopolymer Concrete Waste". In: *Mater. Res.* AHEAD (2017).

[88] C. David. "Thermal degradation of polymers". In: *Comprehensive chemical kinetics*. Vol. 14. Elsevier, 1975, pp. 1–173.

[89] M. A. Villetti et al. "Thermal degradation of natural polymers". In: *J. Therm. Anal. Calorim* 67.2 (2002), pp. 295–303.

# Part III

# FTIR SPECTROSCOPY FOR BREAST CANCER PROGNOSIS

CHAPTER 8

# Introduction

Breast cancer is one of the leading causes of death of females worldwide [1, 2]. Figure 8.1 shows some key statistics about breast cancer in the UK. One in eight women in the UK will be diagnosed with breast cancer at some point in their lifetime, and one in five cases of breast cancer are in woman under 50 [3]. The outlook at the diagnosis of breast cancer differs widely; from death within one year, to the patients being alive and well 30 years after treatment [4]. It is well established that appropriate methods for early and improved prognosis can lead to longer patient survival.

For a typical traditional prognosis procedure of breast cancer, tissue samples are collected from the organ during biopsies (taking a small sample of cells or tissue from breast) or surgical excisions. Extracted tissue samples are sectioned, stained, embedded and fixed onto a glass slide for review. One or more trained histopathologists will observe the processed tissue under an optical microscope and produce a prognosis report based on parameters such as tumour size, stage of the disease, and tumour grade [4]. This cancer histopathology is

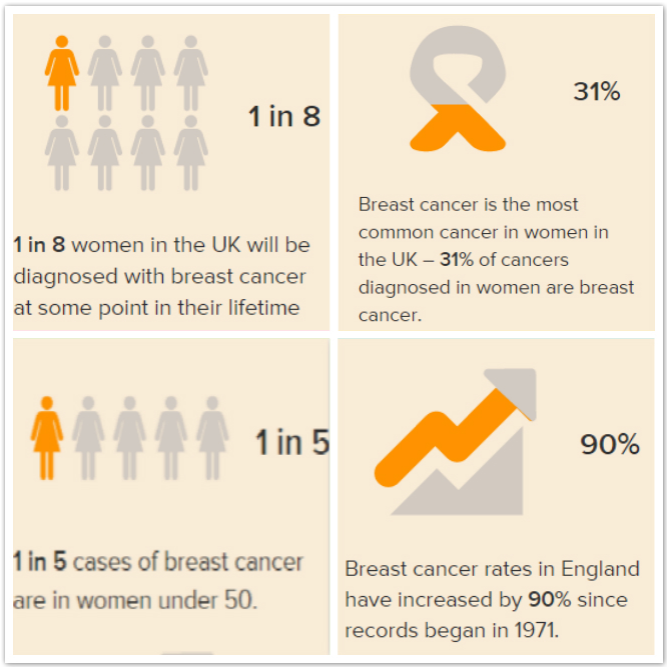



**Figure 8.1:** Breast cancer in the UK. Breast cancer is the most common cancer in women in the UK, 31% of cancers diagnosed in women are breast cancer. Breast cancer rates in England have increased by 95% since records began in 1971. One in eight women in the UK will be diagnosed with breast cancer at some point in their lifetime, and one in five cases of breast cancer are in woman under 50. Reproduced from [3].

time-consuming, highly human dependent and can be inaccurate.

The use of the FTIR spectroscopy for bio-material analysis, which is termed FTIR biospectroscopy, is a rapidly developing research area. FTIR biospectroscopy is a well-established method, that has been investigated through a variety of different research studies [5–11]. These previous results have shown the potential application of FTIR spectroscopy cancer pathology such as breast [12], colon [13], lung [14] and prostate [15], but mainly focusing on diagnosis rather than prognosis.

In this work, we evaluate the prognosis ability of FTIR spectroscopy for identi-

fying different histological grades of breast cancer. Different pre-processing and quality control methods are necessary and we have applied such approaches to each spectrum to correct or minimise spectral problems. Pixels in the background and detector edge area have been distinguished and cut off to further increase the representativeness of the data. Peak symmetric analysis, PCA and standard score (zscore) analysis were used to reduce the dimensions of the data. To combine spectral and spatial information, density estimation (probability density function) was adopted as an analysis tool.

## 8.1 The use of FTIR spectroscopy for cancer histology

The understanding of biochemical changes of cancer is hindered largely because of lacking tools to analyse cell composition. Haematoxylin and eosin (H&E) stained, is the most-used method for breast cancer tissue histopathological examination, and can be used to distinguish nuclei and protein. However, limited molecular information, which is highly relevant to cancer developments, can be extracted. Molecular vibrations are infrared (IR) active and quantitatively measurable by IR spectroscopy [16], providing a non-destructive, label-free tool for molecular composition studying. Figure 8.2 is a typical human breast cancer IR spectrum. Different spectral regions are associated with different molecular vibrations [17]. As highlighted in green colour (Figure 8.2), the Amide I and II wavenumber region in the FTIR spectrum is directly linked with protein secondary structure information. A large and growing body of literature has investigated the application of IR spectroscopy to investigate bio-

samples. There have been many studies in cytological, histological, microbial studies, etc. [18–23]. Different type of samples (e.g., fixed cytology and tissue sections [24], live cells [25] or biofluids [26]) can be analysed.

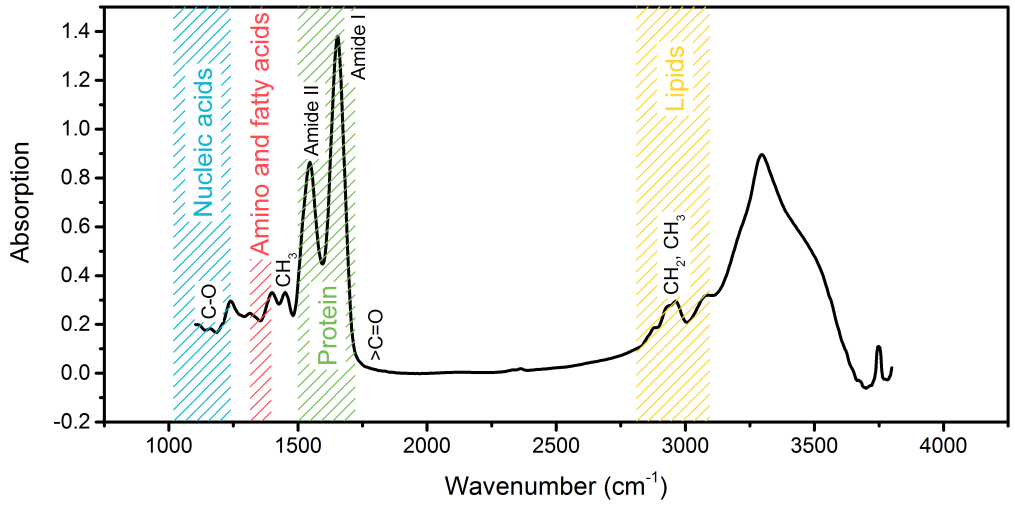



**Figure 8.2:** Typical breast cancer tissue FTIR spectrum showing biomolecular peak assignments from 1000-3800 $cm^{-1}$, where 1085-1241 $cm^{-1}$ = nucleic acid vibrations, 1300-1400 $cm^{-1}$ = amino acid side chains and fatty acids vibrations, Amide I, II = protein vibrations and 2800-3100 $cm^{-1}$ = lipids vibrations. The spectrum is a mean-spectrum of 512 $\times$ 512 transmission imaging from a Grade one human breast cancer tissue. Reproduced from [27].

This section discusses the essential setting parameter (sampling mode, spectral and spatial resolution) of FTIR spectroscopy for cancer histology. Detailed reviews concerning theory, experimental techniques and application examples of FTIR spectroscopy can be found [28–39].

There have been many studies using FTIR for cancer diagnosis with a few examining cancer prognosis. These are highlighted during the text as we describe our approaches to the breast cancer samples we have examined.

### 8.1.1 FTIR spectroscopy sampling mode

There are four major IR spectroscopy sampling modes: transmission, transflection, reflection and attenuated total reflectance (ATR) [40]. Each of them has its own benefits and drawbacks for different samples. Pure reflection measurements are limited to polished samples with a reflective surface. Transmission and transflection have been applied to a variety of research. They are the most popular modes in spectroscopy. Scattering is a problem more related to transmission and transflection modes. It can cause large distortion to the spectra. ATR is usually used in special cases of either thick or highly reflective samples.

Kochan et al. [41] found that both transmission and transflection measurement modes are equally capable of discriminating normal from cancerous tissue in canine liver cancer. Cao et al. [42] indicated that proper pre-processing methods could minimise the classification results difference on transmission and transflection IR spectroscopy modes.

In transmission mode, the sample is placed on an IR transparent surface, such as barium fluoride ($BaF_2$) or calcium fluoride ($CaF_2$). By passing a beam (generated by thermal, synchrotron or quantum-cascade) through the sample and measuring the spectral intensities received by the detector or detectors placed on the other side of the sample, the absorbency of the sample could be obtained.

Barium or calcium fluoride is popularly used as the substrate for tissue to collect FTIR spectroscopy data in transmission mode. These infrared transmissive materials are expensive and fragile. Focusing on the high wavenumber range (2500-3800 $cm^{-1}$), Bassan et al. [43] demonstrated the possibility of using glass

substrates. Support vector machine (SVM) based spectral biomarkers construction followed by a random forest supervised classification, four cell-types prediction was achieved using tissue on glass substrates.

Focusing on the wavenumber region 3125-3700 $cm^{-1}$, Pilling et al. [44] on applying FTIR histopathology using H&E stained glass slides is a major step forward in making FTIR histopathology an economic practical clinic application. Principal components analysis (PCA) was the main noise reduction method before applying random forest classification to construct a classifier to differentiate between epithelium, stroma, blood and concretion. Decomposing spectra into principal components, retaining the lower order PC's and recombining the data set can effectively improve the signal-to-noise. Good improvements in spectral signal-to-noise were observed when utilising PCA based noise reduction with 15 PC's [45].

### 8.1.2 FTIR spectroscopy spectral resolution

Spectral resolution needs to be carefully selected for analysis of cancer samples. A lower spectral resolution could reduce the data acquisition time which heavily influences the clinical or industrial application of FTIR spectroscopy. Low spectral resolution, however, could potentially cause the loss of essential information and choosing the appropriate spectral resolution is a trade-off between acquisition time and information level in the spectra. Previous simulation studies [46–48] have demonstrated that a coarser spectral resolution of 16 $cm^{-1}$ is usually sufficient. There is no doubt that the higher the spectral resolution may provide potentially more information that we gain from each of the spectra at

low resolution.

### 8.1.3 FTIR transmission imaging spatial resolution

The type of detectors being used means the acquisition of spectra could be described as mapping or imaging [31]. The use of the single-element detector (mapping) allows for individual point high signal-to-noise ratio (SNR) spectra being collected across a given sample. Although time-consuming, point spectra often result in high-quality spectra. In contrast to the mapping technique, instruments such as focal plane arrays (FPA) and linear array detectors allow imaging of the sample tissue using spatially arranged detectors. A larger area of interest can be investigated using FTIR imaging technique. Combined with suitable optics, the imaging approach produces good signal-to-noise ratio and spatial resolution FTIR spectroscopy data.

The imaging IR spectroscopy technique is so called because the process of collecting the spectra is similar to take pictures using a camera. In the process of generating a "picture", the multi-sensor detector inside of the camera captures the light information of all the pixels in one go. The spatial resolution of the image depends on the density of sensors on the detector. The diffraction limited resolution and the practical resolution of transmission and ATR mode of FTIR spectroscopy are given in Table 8.1.

FTIR spectroscopy spatial resolution makes the accuracy level a limitation to further apply the FTIR imaging for a medical purpose. Currently, the most popular used pixel size in FTIR imaging is 5.5 $\mu m^2$ [44, 49]. Replacing the

**Table 8.1:** Abbe diffraction limit

| Sampling Method | Diffraction limited resolution | Practical resolution limit |
|---|---|---|
| Transmission | $2\lambda$ | $\sim$10-30 $\mu$m |
| ATR | $0.5\lambda$ | $\sim$3-10 $\mu$m |

low-brightness thermal sources with the synchrotron sources can achieve better signal-to-noise at the high spatial resolution. This technique, however, is challenging to use with wide-field imaging [50].

Cancer tissue is a mixture of cells with different histologic classes. Different cell classes have different shapes, and the average size between classes varies. Even cells within the same class can range from several to dozens of microns [51, 52]. The various sizes of tissue components require high-level spatial resolution to achieve high representativeness of the spectra. Figure 8.3 is the imaging of prostate tissue under different spatial resolution setting. Compared with the H&E image, the 5.5 $\mu m^2$ spatial resolution can provide limited tissue structure information.

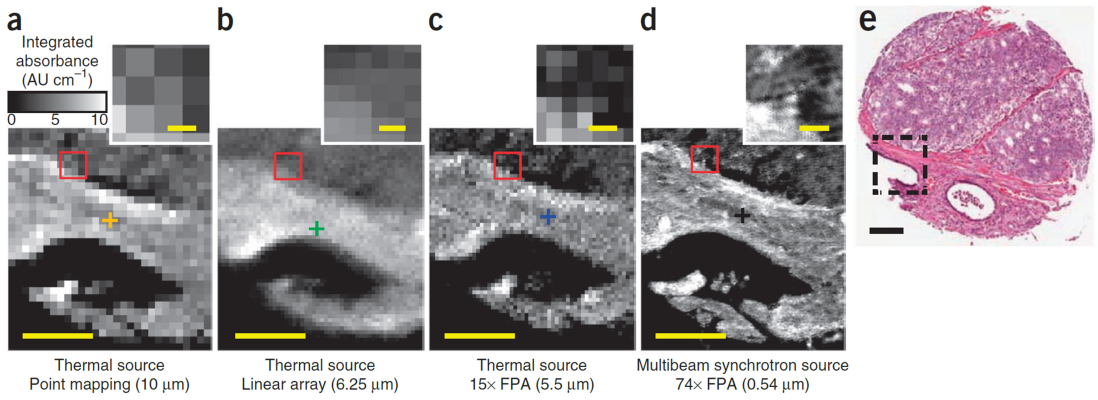



**Figure 8.3:** Pseudo-colour images from various FTIR spectroscopy systems. (a-d) The same cancerous prostate tissue section (area, 280 $\mu$m $\times$ 310 $\mu$m) measured with different instruments, using the integrated absorbance of the CH-stretching region (2800-3000 cm$^{-1}$), without dyes or stains. All images were processed identically (baseline correction only) and used the same colour scale. Scale bars, 100 $\mu$m and in insets, 10 $\mu$m. (a) Images acquired with a conventional table-top system (PerkinElmer Spotlight) equipped with a thermal source in raster-scanning mode (10 $\mu$m $\times$ 10 $\mu$m; (b) and linear array mode (6.25 $\mu$m $\times$ 6.25 $\mu$m; (c) with an FTIR imaging system (Varian Stingray) equipped with a 64 pixel $\times$ 64 pixel FPA (5.5 $\times$ 5.5 $\mu$m per pixel at the sample plane; (d) and with multibeam synchrotron-based imaging system (pixel size, 0.54 $\mu$m $\times$ 0.54 $\mu$m. (e) Hematoxylin and eosin (H&E)-stained prostate tissue (diameter, 0.75 mm). Scale bar, 100 $\mu$m. Dashed box specifies the corresponding area of a serial, unstained section from which generated images in a-d. Reproduced from [50].

## 8.2 FTIR spectroscopy on breast cancer histology

### 8.2.1 Nottingham Prognostic Index

In 1982, based on a study on 387 patients, researchers at Nottingham City Hospital created a breast cancer prognosis index named the Nottingham Prognostic Index (NPI) [53]. Since then it has become a widely accepted as the major prognosis index for breast cancer. The factors used to give a significant correlation with prognosis were tumour size (measured pathologically), histological lymph node stage and histological grade [53]. Not only can the disease be recognised, but also unusual cases, such as deficiencies in diagnostic quality could be detected through manual examination of the tissue [48].

To measure the Nottingham Prognostic Index, excised material is initially formalin fixed to prevent its degradation, and subsequently prepared onto glass substrates for optical microscopic analysis. The identifications of cellular and extracellular components within the tissues and cells are enhanced by the addition of dyes, which stain different component different colours [27].

Hematoxylin and Eosin stain (H&E stained or HE stained) is one of the principal stains used [54, 55]. Hematoxylin has a deep blue-purple colour and stains nucleic acids by a complex, unknown reaction. Eosin is pink, and stains proteins [56]. In a typical tissue, nuclei are stained blue, whereas the cytoplasm and extracellular matrix have various degrees of pink staining. Well-fixed cells show considerable inter-nuclear details. The stain discloses abundant structural information, with specific functional implications. A typical H&E stained tissue is shown in Figure 8.4, with the individual tissue structure identified.

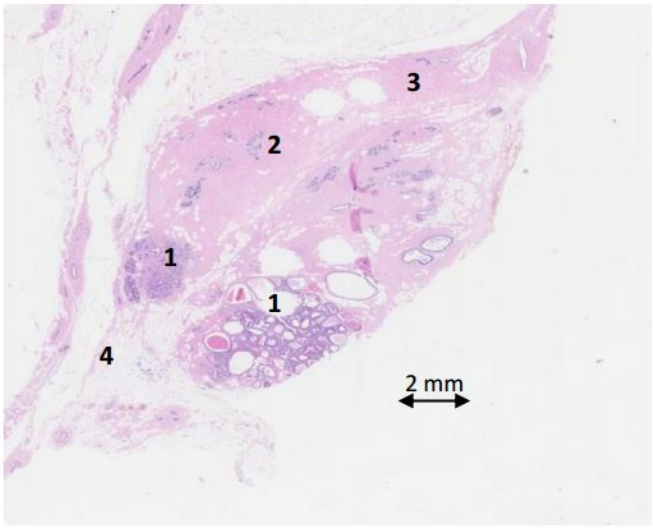



**Figure 8.4:** H&E stained parallel section of grade one breast cancer tissue showing (1) tumour (2) normal cells, (3) stroma and (4) fatty tissue. Reproduced from [57].

Using the coefficients of significance, a simple numerical prognostic index (PI) can be devised as follows:

$$PI = 0.2 \times tumoursize(diameter\ in\ cm) + lymphnodestage(1-3) + histologicalgrade(1-3). \quad (8.2.1)$$

The possible values of lymph node stage are:

1. no nodes affected (1 point)
2. up to 3 nodes are affected (2 points)
3. more than three nodes are affected (3 points)

The possible values of tumour grade are:

1. G-I: less aggressive appearance of tumour (1 point)

2. G-II: intermediate appearance of tumour (2 points)

3. G-III: more aggressive appearance of tumour (3 points)

The prognosis on this index worsens as the prognostic index increases, and by using cut-off points of 3.4 and 5.4; patients may be stratified into good (grade one), moderate (grade two) and poor prognostic (grade three) groups having an annual mortality rate (units of deaths per 100 individuals per year) of 3%, 7% and 30%, respectively [58].

Patients often wait significant periods to get the results [59] and the reproducibility of NPI can be problematic because of the level of human involvement in the prognosis. Only 50 % to 85 % of cases that were investigated by different pathologists have the same results [60]. It is still not regarded as an important procedure in routine breast cancer diagnostics in some units because of apparent inter and intra-observer variability [61, 62].

### 8.2.2 FTIR spectroscopy breast cancer diagnosis/prognosis

Yun Xiang Ci et al. reported the applying of FTIR spectroscopy on breast cancer diagnosis [9]. Small amount of the frozen tissue was minced, dispersed and centrifuged. The cell suspension was then dried on an IR transparent slide to form a thin film for IR spectroscopy study. The spectra difference indicated an increase in nucleic acid in carcinoma tissue, as well as an increase of collagen in fibroadenoma tissue. Those results are consistent with the histopathological examination results.

An attempt of understanding the structural changes in the breast cancer tissue

during different steps in the development of tumour using FTIR spectroscopy is made by Eckel et al. [8]. Breast tissue is used directly in this study. The diagnosis result is based on another stained tissue from the same surgical breast resection. The result IR spectra characteristic indicated that protein/collagen ratio is significantly higher in the carcinomatous tissue.

In 2002, Fabian et al. reported the obtaining of IR spectra of breast tumour tissue sections using a microscope equipped with a Focal Plane Array (FPA) detector with a spatial resolution near the diffraction limit [63]. The research reveals a high-sensitivity IR spectroscopy approach to study tissue biochemistry changes. The data demonstrated the need for high spatial resolution at the level of individual cells.

In 2006, the same group reported the obtaining of IR spectra of benign and malignant lesions in breast tissue sections [64]. Conventional point-by-point mapping, 16-element linear array detector imaging and 64 $\times$ 64 FPA detector imaging were used for FTIR spectroscopy data acquisition. The work provides further insights into the spectral 'averaging' problem prevented us from obtaining 'pure' spectra of histological class, when the spectra were acquired using aperture diameters of 30 - 40 $\mu$m. An artificial neural networks (ANN) model was trained to perform automated classifier to separate fibroadenoma, ductal carcinoma in situ, connective tissue, and adipose tissue, based on IR spectra data.

Using a 64 $\times$ 64 Focal Plane Array (FPA) detector, Kumar et al. applied FTIR imaging technique on tissues obtained from formalin fixed, paraffin embedded (FPPE) tumour blocks to the extracellular matrix. FTIR spectral differences

were reported when examining the extracellular matrix close to and far from carcinoma. Major changes were observed in the relative intensities of the collagen bands at 1640 and 1630 $cm^{-1}$ [65].

Pounder et al. [7] was an early study trying to examine the potential to combine the spectral and spatial information for breast cancer diagnosis. Each pixel was first labelled as stroma or epithelium using spectral recognition at the single pixel level; subsequently, epithelium pixels are labelled as cancer or normal by spatial polling based upon epithelium content and distribution. Results indicated that the protocol is highly accurate under a variety of conditions.

In a more recent study by Berisha et al. [66], convolutional neural networks (CNNs) are applied to data from breast cancer tissue microarrays to identify six major cellular and acellular constituents of tissue, namely adipocytes, blood, collagen, epithelium, necrosis, and myofibroblasts. Experimental results show that the use of spatial information in addition to the spectral information brings significant improvements in the classifier performance and allows classification of cellular subtypes.

The studying of breast cancer using FTIR spectroscopy has been investigated at Nottingham with Bird et al. [67] reporting the using of FTIR mapping technique to analyse frozen and deparaffinised tissue for diagnosis purpose. The result spectra have the pixel of 25 $\mu$m $\times$ 25 $\mu$m in size. An ANN-based diagnosis algorithm was developed to perform classification of benign and malignant tissues composed within different lymph nodes.

Wang [68] focuses on the development of fuzzy clustering techniques to invest-

igate the use of infrared spectroscopy as a diagnostic probe for the identification of the early stages of cancer. Aiming to automatically identify the different types of tissue present within any given spectral dataset, they developed a simulated annealing based clustering algorithm and a new technique to merge the clusters with the same biochemical characteristics.

Stapleton [57] collected FTIR images (mapping) from eight breast cancer tissue sections (2 Grade one, 3 Grade two and 3 Grade three). Each of the FTIR datasets was subjected to Principal Component Analysis (PCA) to explore the patterns in the data. Two methods were tried to construct false colour images, PCA/Fuzzy c-means clustering and Multivariate Curve Resolution. The same approach was adopted to a data set with imaging the tissue sections with a focal plane array detector with pixel size 5.5 $\mu$m. In the future work section of Stapleton's thesis, they reported another 30 tissues, 10 of each grade, have been prepared and FTIR data has been recorded. This is the main data set have been used for the analysis in this thesis. A more detailed description of the data set is given in Section 8.4.2 on page 192.

Naqvi [69] reported a novel model based on Type-II fuzzy logic for breast cancer FTIR spectral data to help clinicians classify breast cancer grades. Results indicate that the model was able to successfully construct prototype fuzzy sets for the data set, and provide in-depth information regarding the complexities of the data set as well as helping in the classification of the data.

The novelty of the research reported in this thesis lies with the introducing of a data processing method which has fewer requirements at the accuracy level of the spectra library. Only observable spectra differences were used to segment

different histologic classes. Combining with the dimension reduction methods in Chapter 10, we are able to produce highly accurate breast cancer prognosis. The statistic methods introduced in the study give more rooms for errors, noise and mixture pixels in the segmentation step, which fits state of the art FTIR spectroscopy machine spatial resolution. The validation in Chapter 11 shown that the algorithm developed has good generalisation ability and has the potential to be widely used on other biological/medical FTIR data.

Table 8.2 is the list of methods and example applications given in a review conducted by Trevisan et al. [70]. Those methods highlighted in orange are the one reported in this thesis.

**Table 8.2:** List of FTIR spectroscopy data analysis methods with methods used in this thesis highlighted. Reproduced from [70].

| Stage | | Name of method |
|---|---|---|
| Quality Control (QC) | | Water vapour check |
| | | Signal-to-noise test |
| | | Thickness test |
| | | Maximum absorbance threshold |
| | | Clustering |
| Pre-processing (PP) | De-noising | Savitzky-Golay Smoothing |
| | | Wavelet de-noising |
| | Removal of physical phenomena | 1st Differentiation |
| | | 2nd Differentiation |
| | | Rubberband baseline correction |
| | | Manual baseline correction |
| | Fitting methods | Polynomial |
| | | Extended Multiplicative Signal Correction (EMSC) |
| | | Resonant Mie Scattering Correction (RMieSC) |
| | Normalisation | To the Amide I peak |
| | | To the Amide II peak |
| | | Vector normalisation |
| | | Area normalisation |

**Table 8.2:** List of FTIR spectroscopy data analysis methods continued from previous page. Reproduced from [70].

| Stage | | | Name of method |
|---|---|---|---|
| Feature Extraction (FE) | Feature Construction (FC) | Linear | Principal Component Analysis (PCA) |
| | | | Partial Least Squares (PLS) |
| | | | Linear Discriminant Analysis (LDA) |
| | | | PCA-LDA |
| | | Others | Band fitting |
| | | | Wavelet transform |
| | | | Peak picking |
| | | | Relative distance plane |
| | Feature Selection (FS) | Wrapper methods | Forward Feature Selection (FFS) |
| | | | Genetic Algorithm (GA)-based |
| | | | GA_ORS |
| | | Filter methods | Kruskal-Wallis test |
| | | | Neurodeveloper's COVAR |
| | | Embedded methods | Inside eClass |
| | | | Random forest |
| Clustering | | | k-Means |
| | | | Fuzzy c-means |
| | | | HCA |
| Classification | | | Linear Discriminant Classifier (LDC) |
| | | | PLS-Discriminant Analysis (PLS-DA) |
| | | | Support Vector Machine (SVM) |
| | | | Artificial Neural Network (ANN) |
| | | | k-Nearest Neighbours (k-NN) |
| | | | Hierarchical ANNs |
| | | | Classifier ensembles |
| | | | Incremental/evolving classifiers |

## 8.3 FTIR spectroscopy Data Analysis

FTIR spectroscopy data provides multivariate information based on chemical-specific IR spectra. By passing the FTIR spectroscopy data through a variety of data analysis algorithms, we can re-construct tissue images or cell architectures. For a complex sample, such as human being tissue, FTIR spectra are a superposition of all the spectra of the individual biochemical components [71]. Those complex spectral signatures are decoded through the data analysis steps. The fingerprint spectra constructed data or image is a reliable way to understand the health statues of the sample tissue. Despite the increasing popularity of the FTIR bio-spectroscopy field, challenges relating to sample preparation, instrumentation and data handing still remain.

There are two analysis aims in the bio-spectroscopy field: imaging and diagnosis. Imaging derives an image of the tissue architecture expressing the underlying biochemistry from data for further research analysis purpose. Diagnosis is based on supervised or unsupervised learning approach to classify or clustering the spectra in the aim of histopathology. Different work-flow is used to achieve different goals of analysis.

Figure 8.5 is the framework for exploratory and diagnostic using of breast cancer FTIR spectroscopy data. It was first reported by Trevisan et al. [70]. Further improvements were made based on recent literature [39, 44, 72]. The exploratory technique tends to be qualitative and provide graphical visualisation of data, whereas, the diagnostic focuses on quantitative classification and decision making. By using IR spectroscopy either as an image tool or by classifying

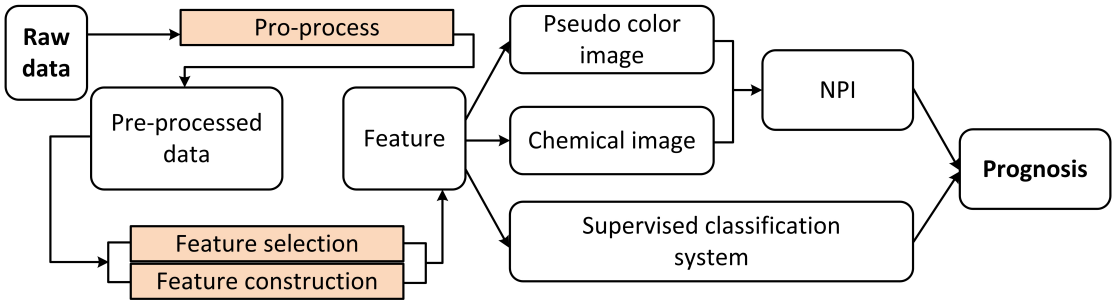



**Figure 8.5:** FTIR biospectroscopy data analysis work-flow for exploratory and diagnostic. For chemical imaging, extracted features are used to decide the colour of one pixel. For pseudo-colour imaging, clustering is commonly used to assign colour to a group of pixels. For diagnostic, supervised learning is applied to extracted feature to build a diagnostic system to make predictions on new data sets. NPI: Nottingham Prognosis Index.

spectral categories, it has been possible to distinguish between benign and malignant tumours in tissue samples of the breast [12], colon [13], lung [14] and prostate [73].

In a paper, published by Baker et al. [39], they brought together some of the leaders in the FTIR biospectroscopy field and produced a manuscript to standardise the methods and procedures of applying FTIR spectroscopy to a variety of biological or clinical questions. They compared the supervised classification performance between different combinations of pre-processing, feature extraction and supervised classification methodologies on the blood plasma tissue sample. The result indicated that no single pre-processing, feature extraction or supervised classification method is the absolute best, but a combination of these may be the best solution. Following sections are detailed introductions to each step given in Figure 8.5.

### 8.3.1 Exploratory

In exploratory, data processing is directed mainly towards visualising to reveal tissue structures. After pre-processing, Feature extraction (FE) method (clustering or supervised classification) for image could be used to reduce the dimension of spectra. The scalar value of the newly extracted feature could be used to determine the colour of that pixel to produce a **chemical image** of the tissue. Further unsupervised learning techniques could be applied to the new features to clustering pixels into different groups. One single value is assigned to the pixels that belong to the same cluster. Based on the newly assigned values, **pseudo-colour image** could be built. The resulting images could either be used to penetrate important histopathological features, or they could be used as the further study materials.

In the false-colour image approach, the continuous absorbance value is changed to discrete numbers to represent the colour is given to that pixel. This data processing step excludes a lot of information that is potentially useful for cancer histopathology. In the chemical image, colour of the pixel is assigned based on the real absorbance difference. There is no information lost in this process. However, this approach transferred the biochemical component difference into colour, which is not easy to conduct quantitative analysis on. The chemical image approach is counterproductive [71] as only one intensity value (or the intensity integration of one peak) of a spectrum is used, which is causing the loss of information contained in the rest of the spectra.

An example of using FTIR spectroscopy for imaging comes from the study con-

ducted by Nasse et al. [50]. By combining multiple synchrotron beams with wide-field detection, they produced high-resolution prostate tissue images over the entire mid-infrared spectrum with high chemical sensitivity and fast acquisition speed while maintaining a high-quality signal-to-noise ratio. Figure 8.3 on page 172 is the chemical image result from various FTIR spectroscopy systems.

The exploratory approach is currently a major direction of FTIR imaging studies. Although this study could provide high-contrast images with minimal sample preparation for traditional histopathology use, further computing techniques are needed to truly utilise the potential of IR spectroscopy on cancer diagnosis and prognosis.

**Clustering** is widely used for pattern finding in exploratory approach. It belongs to the unsupervised pattern recognition category. Based on the similarity between samples, clustering could cluster them into different groups. Clustering is also a useful tool for diagnostic approaches. It can help researchers gain a better understanding of the data structure, which will help in classifier design. The most popular clustering method in biospectroscopy data analysis area is: k-mean, fuzzy c-mean, and hierarchical cluster analysis (HCA) [69, 74, 75].

### 8.3.2 Diagnostic

Using FTIR biospectroscopy data for diagnostic purposes is a more complex procedure. This approach aims to create an efficient, robust and highly reliable computer-based diagnostic (or prognostic) system (DS). Supervised classification methods are employed. It requires a more complex framework from data

pre-processing to classification. A typical diagnoses process includes: quality control, data pre-processing, data mining, feature extraction and supervised classification. Based on the nature of supervised learning, different data sets require different pre-processing, feature extraction and supervised learning methodologies. Different combinations of methodologies should be attempted and compared in any diagnostic study.

**Quality control** consists of identification and removal of outliers. Outliers are measurements or samples considered to be wrong, thus unsuitable for use in the data analysis [70]. They could be caused by all kinds of reasons during the sample preparation and spectra acquisition steps. Different reasons have different effects on the spectra. For example, outlier spectra caused by different thickness of the samples have absorption peaks either too high or too low. The mutation of the environment during the acquisition of spectra could cause random absorption peaks that make the spectrum an outlier.

Most of the quality control methods are employed at the first stage of data processing: data pre-processing. Alternative forms of quality control are available during the classification or other stages. 'Refusing to decide' function of a classifier is considered as a quality control method during the classification step. Based on literature review, quality control seems to be underrated, with only a few studies exploring this stage [39].

**Pre-processing** aims to make the FTIR spectra easier to interpret by correcting issues related to spectra data acquisition. It is the most important stage in FTIR biospectroscopy data analysis and can be divided into de-noising, spectral correction, normalisation, derivatisation and other manipulations. Two or

three methods are often combined, e.g., de-noising followed by spectral correction and normalisation [39]. Different options are available for each of the pre-processing steps. The combinations of these steps may be more or less preferred than others, depending on the sample type, noise level, instruments setup, aim of analysis, classification performance and personal preference. The most popular pre-processing sequence is: de-noising, spectral correction, differentiation, normalisation and other manipulations. It is worth noting, none of the steps mentioned above is mandatory. The optimal pre-processing method or sequence to apply is still under discussion [76–78]. There is still no universal best approach existing for all samples. Several attempts have been made to standardise pre-processing methods [70]. The use of a machine learning genetic algorithm (GA) is a potential source of pre-processing standardisation, allows maximum flexibility for end users [76].

**Feature extraction (FE)** processes the IR spectra to form new features from the original features. In classification or pattern recognition, the 'curse of dimensionality' is often mentioned. This refers to the difficulties to train a classifier in high-dimensional spaces, when models can easily over-fit or remain undertrained [79, 80]. Feature extraction aims to produce a smaller number of variables that are more informative for further classification. Feature extraction has an essential role in both imaging and diagnosis. For imaging, feature extraction could generate a single value for each spectrum, which defines the colour of that pixel in a pseudo-colour image. For prognosis, the feature extraction could reduce the input dimensions of the supervised learning system while incorporating physical knowledge. That would help to avoid over-fitting or

under-training.

One subgroup of feature extraction is feature selection (FS). Based on biochemical understanding of the problem, the features that are highly relevant to the results are selected from the original feature set. This approach is particularly interesting as it can confer biochemical knowledge/information to the classification system. Another subgroup of feature extraction is feature construction (FC). New relevant features are constructed from the original features. Popular feature construction methods for biospectroscopy data analysis include calculating the ratios between wavenumber absorbance intensities [81], area under a sub-region of the spectrum and performing principal component analysis (PCA).

**Supervised classification** is the segmenting the sample domain into groups based on sample input-output pairs. The promise of clustering is that no prior information is fed to the methods for classification. Hence, finding clusters in which intra-cluster variation is smaller than inter-cluster variation is always a problem. For supervised learning, each sample is assigned a priori, which is known as the true value. The data set can then be split into a training set and a test set. The training set is used as teaching information to build models.

Different supervised classification algorithm could be applied when training the model. Support vector machine (SVM) [82], random forest [83], artificial neural networks (ANN) [84], linear discriminant classifier (LDC) [85] and Bayesian inference-base methods [86] are those commonly used methods in the biospectroscopy area. Before the model could be applied to predict the classes of a data set that do not have prior, the test set needs to be applied to test the perdition

ability of the model. Only a model that has a low error rate (the number of incorrect guesses divided by the total number of guesses) could be used to predict. Except for the accuracy, there are still other criteria for choosing classifier: easy to train, computational time, software ability, etc. If two different classifiers were performing equally well on an independent test set, the simpler one should be chosen for a better generalisation ability [79].

The optimal size of a training data set has been under investigated to date, but it has been suggested that it may be problem dependent [87]. A general guideline for applying supervised techniques is to have the number of sample 5-10 times bigger than variables [88]. Cross-validation could be used when training the model using a small training data set.

### 8.3.3 Software package

The instrumentation software, coming together with the FTIR spectroscopy machine provides a number of pre-processing and sometimes more advanced data analysis options. Various data analysis software program packages, ranging from basic de-noising methods to targeting specific data analysis tasks exist. MATLAB is the most popular programming language environment due to its special ability to process metrics. Customised software could be written for specific tasks. Python is another programming language that is becoming increasingly popular in the FTIR spectroscopy data analysis field. It has advantages in using and creating open source packages. All the data analysis in this thesis was performed using MATLAB R2018a (The MathWorks Inc., Natick, MA).

## 8.4 Knowledge discovery cycles and data description

"*Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.*"

– Henry Nix, 1990

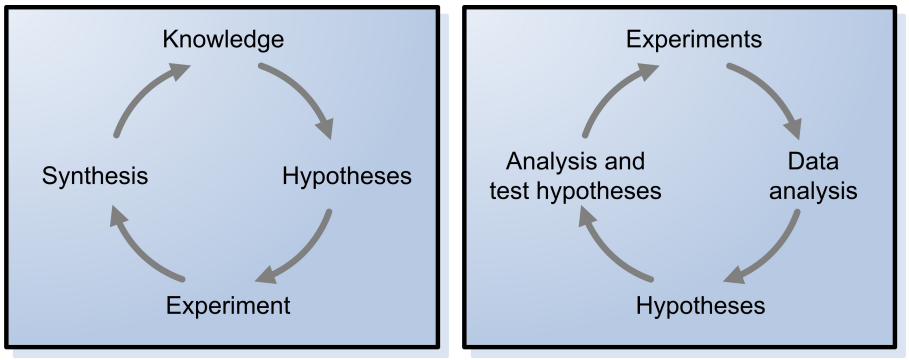
### 8.4.1 Knowledge discovery cycles

The generation of hypotheses, the gathering of data and the evaluation of hypotheses to transfer them into knowledge is sometimes referred to as the 'cycle of knowledge' [89].

**The knowledge-based cycle** is popularly used before the available of modern computing facilities. It is also referred as the traditional cycle. In that cycle, we start with some basic knowledge about the problem domain. Hypotheses are generated based on those prior-knowledge which will be tested by the specifically designed experiments. By 'deductive reasoning', based on the experiment observation, new knowledge about the problem domain is generated. The cycle then repeats itself.

**The data-driven knowledge discovery cycle** is applied in the early stage of research in which our knowledge about the problem is minute. Large amount of data can be easily gathered. Based on this great many observations, algorithms or inductive reasoning is applied to generate hypotheses. Strong and robust

algorithms are needed to abstract knowledge from data in this case. As with any purely inductive method, there are no axioms and so the hypotheses that evolve cannot be proved correct. However, they greatly narrowed the search space of possibilities, meanwhile increase the understanding of the problems. The knowledge-based cycle is given in Figure 8.6a. The data-driven cycle is given in Figure 8.6b.




**(a)** Hypothetico-deductive reductionist approach

**(b)** Inductive approach

**Figure 8.6:** The cycle of knowledge discovery. (a) Hypothetico-deductive reductionist approach, in which background knowledge is needed to generate a hypothesis. (b) Inductive approach, in which the hypothesis is generated from the data. It is a data-driven approach, which is highly relying on data mining. Reproduced from [90].

Inductive/data-driven approach was used in the breast cancer work reported in Part III Applying FTIR spectroscopy on Breast Cancer Prognosis. Algorithms were developed based on the interpretation of the breast cancer tissue IR spectra. Hypotheses were generated based on patterns revealed by those algorithms. Data was manipulated and presented in various ways. Based on the observations of the data, attempts of explaining cancer developments based on the data pattern were made. Due to the limited understanding of bio-molecular

developments of cancer, it is difficult to have direct observations that are consistent or inconsistent with the hypotheses.

Hypothetico-deduction approach was used in Part II Applying FTIR Imaging to Address Challenges in Plastic Recycling. The algorithms and visualisation method developed in Part III was further applied and tested using a much simpler sample. Studying polymer using IR spectroscopy is a well-developed research field. The data patterns observed can be well supported by the IR absorption knowledge of plastic. The confirmation of the hypotheses provided further confidence in the analysis results in Part III.

### 8.4.2 Data description

For FTIR spectroscopy biomedical or bio-spectroscopy study, typical pathology specimen, formalin-fixed, paraffin-embedded (FFPE), is widely used [39].

Using a microtome, two 6 $\mu$m thickness contiguous sections were cut from an FFPE tissue block, as illustrated in Figure 8.7. One of the two tissues was mounted onto calcium fluoride ($CaF_2$) for transmission mode FTIR imaging. Another adjacent section was H&E stained for histopathology evaluation. The results of the evaluation were taken as the true value. Taking contiguous sections ensure that the tissue for histopathology and FTIR spectroscopy data collection has the most similarity. Removal of paraffin is essential, since it has strong IR absorbance between 2800-3000 $cm^{-1}$. A standard de-waxing method was employed to remove paraffin from the tissue before any analysis can be conducted. The procedure of data collection is detailed discussed by Stapleton [57].

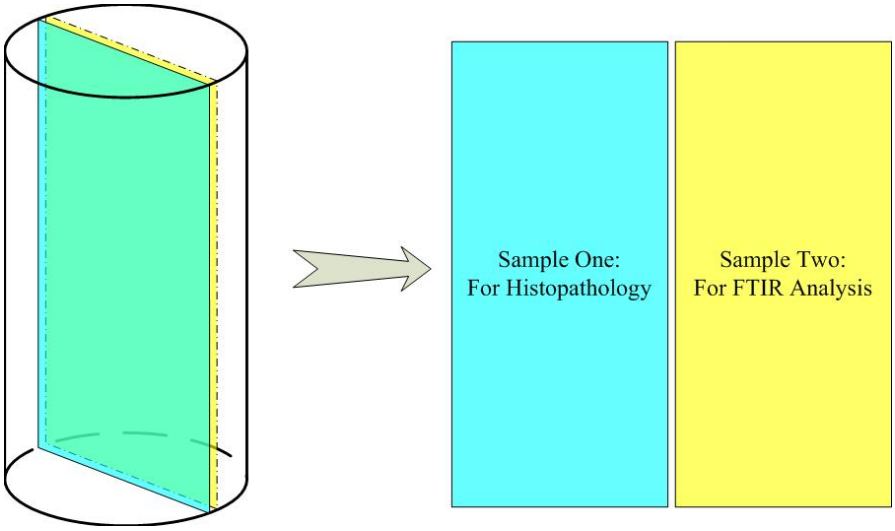



**Figure 8.7:** Collection of tissue. All tissues were collected during routine biopsies (taking a small sample of cells or tissue from breast) of patients, and they all have a full histological diagnosis and a complete clinical history. The tissue columns getting from biopsies then were sliced along the axial direction. The tissue used for FTIR imaging, and the tissue used for H&E stain are neighbouring sections.

30 tissues, 10 of each grade, have been prepared and FTIR imaging data have been recorded. All the data were obtained on transmission mode with a 3.8569 $cm^{-1}$ step size (7.7138 $cm^{-1}$ spectral resolution), 5.5 $\mu$m spatial resolution. 128 $\times$ 128 focal plane arrays detectors are used. For each sample, 4 $\times$ 4 areas were imaged. Therefore, 512 $\times$ 512 spectra were collected per sample. All data sets are truncated to 1000-4000 $cm^{-1}$ for easy data handing and storage. Infrared spectra from each tissue sample were stitched together in MATLAB to form a single (512 $\times$ 512 $\times$ 700) data cube, consisting of 262144 spectra with 700 data points each.

We use a numerical research identifier for each of the 30 tissues. To denote the sample tissue cancer status, 1-10 are assigned to the 10 samples of grade one,

**Table 8.3:** 30 sample dataset summary

| NO. | Research Identifier | Tumour Grade | NO. | Research Identifier | Tumour Grade | NO. | Research Identifier | Tumour Grade |
|---|---|---|---|---|---|---|---|---|
| 1 | ERS 091 | 1 | 11 | ERS 027 | 2 | 21 | ERS 009 | 3 |
| 2 | MI 4197 | 1 | 12 | MI 4258 | 2 | 22 | ERS 013 | 3 |
| 3 | MI 4502 | 1 | 13 | MI 4373 | 2 | 23 | ERS 016 | 3 |
| 4 | MI 4531 | 1 | 14 | MI 4569 | 2 | 24 | ERS 018 | 3 |
| 5 | MI 4572 | 1 | 15 | MI 4758 | 2 | 25 | ERS 024 | 3 |
| 6 | MI 4690 | 1 | 16 | ERS 069 | 2 | 26 | ERS 076 | 3 |
| 7 | MI 5061 | 1 | 17 | ERS 075 | 2 | 27 | ERS 102 | 3 |
| 8 | MI 5103 | 1 | 18 | ERS 109 | 2 | 28 | MI 4108 | 3 |
| 9 | MI 5267 | 1 | 19 | ERS 111 | 2 | 29 | MI 4117 | 3 |
| 10 | ERS 014 | 1 | 20 | MI 4107 | 2 | 30 | MI 4121 | 3 |

11-20 are assigned to the 10 samples of grade two, while 21-30 are assigned to the 10 samples of grade three. Details of the 30 tissue sample can be found in Table 8.3, in which the research identifier for tissue bank, the numerical research identifier ('NO' in Table 8.3) used in this thesis and the grading information of each tissue are given.

CHAPTER 9

# Pre-processing & segmentation

Human breast tumours are histologically complex tissues, containing a variety of cell types in addition to the carcinoma cells [91]. Histologic segmentation is necessary in order to provide automatic prognosis. In the pages that follow, different pre-processing algorithms were discussed and tested. The best-performed algorithms were applied to breast cancer tissue spectra to improve robustness and accuracy. A spectra library containing stroma, normal, carcinoma cell spectra was built. Background empty pixels and pixels lie on the edges of the FPA detector were identified and cut off.

## 9.1 Pre-processing

Pre-processing is normally the first step taken in FTIR spectroscopy data analysis to correct baseline slope, different sample thickness or concentrations, scattering and noise problems in raw FTIR spectra. It is a very delicate procedure, in which, special cares are needed in this step to avoid miss-interpretation of the spectra.

RMieS-EMSC [92, 93] pre-processing method was utilised to correct the baseline problem caused by Mie Scattering. The key parameter setting for the algorithm is given in Table 9.1. The result indicates the algorithm with the parameter setting can effectively remove the baseline of the spectra [57].

**Table 9.1:** Parameters used for Mie scattering correction algorithm

| Parameter | Value |
|---|---|
| Lower wavenumber point | 1000 |
| Upper wavenumber point | 4000 |
| Number of Iterations | 1 |
| Number of principle components | 8 |
| Mie theory option | 2 (RMieS) |

The inter-pixel biochemistry variation makes each pixel (spectrum) contains different levels of noise or thickness variation, which, requires a distinctive level of pre-processing treatments. A high computational power algorithm can overcorrect spectra causing the loss of valuable cancer development information. A weak algorithm, however, under-corrects spectra and introduces a high noise level that can swallow the cancer relevant variation.

As the biochemical component in each pixel is unclear, it is difficult to objectively reason the effects of each pre-processing step. A detailed, knowledge-based, analysis of pre-processing algorithms is made in Section 6.1 based on plastic film sample. The conclusions that we draw are used to supervise the pre-processing steps for breast cancer tissue FTIR spectroscopy data. As discussed in Section 6.1, Savitzky-Golay smoothing, with parameter setting SPAN=9, DE-

GREE=2, is used to remove the background and system noise.

Using a microtome, the thickness of the tissue is well maintained to be 6 $\mu$m across the sample. However, for breast cancer tissue, the chemical compound containing in each pixel is different as the bio-chemical material is not evenly distributed across the sample. Normalising the spectrum to a certain peak makes the assumption the chemical structure represented by that peak can be used as an internal standard, as the value in a normalised spectrum means the absorbance ratio other than the actual absorbance. However, as discussed in Section 6.1, the noise level of the whole spectrum could be heavily influenced by the S/N ratio of the normalisation data point. For different analysis algorithms, normalised or non-normalised data set should be used. For example, PCA, can automatically abstract the major variation between spectra. Using non-normalised spectra, the major variation containing in the data set will be the variation caused by uneven distribution. Normalised spectra are used in the following spectra library Section 9.2.1 for PCA analysis as the major variation wanted is the bio-chemical difference caused by different histological classes. The results indicate that using the normalised spectra can provide better divination between histologic classes. In the zscore analysis Section 10.7, inputting non-normalised spectra gives better results, as the algorithm is taking each wavenumber absorbance as individual population. Using normalised data set, the algorithm is looking at the composition ratio of two structures other than the composition of one structure across the sample.

In conclusion, the selected pre-processing methods and the sequence of applying them are as following.

1. RMieS-EMSC method is used to correct the Mie scattering problem [57] (data set identifier in coding: ZCorr);

2. All spectra were truncated to 1100-3000 $cm^{-1}$ (500 data points for each spectrum), for easy data handling and storage;

3. Savitzky-Golay smoothing algorithm with 'SPAN' equals to 9, 'DEGREE' equals to 2 is used to remove noise caused by random background or system fail (data set identifier in coding: ZCorr_smooth);

4. Amide II peak normalisation was applied to each spectrum. The maximum absorbance between 1543-1551 $cm^{-1}$ of each spectrum was identified. Adjusting the Amide II peak value to 1, by dividing all spectrum by the peak absorbance. (data set identifier in coding: ZCorr_normal);

5. First principle differentiation was applied to each spectrum (data set identifier in coding: D).

Each step is based on the spectra result of the step before, which will make the current result the result of sequentially applying all the previous steps. For instance, first principle differentiation dataset (dataset identifier in coding: D), has the pre-processing steps of: RMieS-EMSC $\rightarrow$ spectra truncation $\rightarrow$ Savitzky-Golay smoothing $\rightarrow$ Amide II peak normalisation $\rightarrow$ first principle differentiation.

In the process of collecting FTIR spectroscopy data, a background spectrum is collected and used to minimise the environmental background absorbance. The choice of the background spectrum is essential. A high absorbance level

background spectrum could cause the real empty spectra to have all negative absorbance. The normalisation for negative spectrum is different from what reported above. The normalisation of those spectra was conducted by dividing the spectrum by 0.1, other than the Amide II peak value itself. This is to guarantee those negative spectra can be identified and cut off by the empty pixel segmentation method.

The optimal pre-processing method or sequence to apply is a subject of discussion, and no universal best approach exists for all samples [39]. In this thesis, for different further analysis algorithm, different data set is used. A clear data description, as well as the reason for using this data set is given before each of the further analysing methods.

## 9.2 Tissue segmentation

In research published by Fernandez et al. [73], they coupled FTIR imaging with tissue microarrays with statistical pattern recognition and demonstrated histopathologic characterisation of prostatic tissues. They examined more than 9.5 million spectra from over 870 samples, and reported a subset of 3 million spectra from 262 samples at a spatial resolution of 6.25 $\mu$m. Individual cell type spectra (normal epithelium, fibrous stroma, mixed stroma, muscle, nerve, lymphocytes, stone, ganglion, endothelium and blood) were compared. Special pattern descriptors that established differences between specific cell types were found and termed as 'metric'. Automated histologic segmentation was achieved with well-defined tests of statistical significance. The detection of

prostatic adenocarcinoma, which conventionally requires a manual analysis of the histomorphology and growth of tumour cell ensembles, was automated by examing epithelial spectra. They successfully differentiated benign from the malignant prostatic epithelium by spectroscopic analyses.

Pilling et al. [44] reported the effects of using coverslipped H&E stained tissue on glass to conduct spectral histopathology. Using a spectral database constructed by histopathologists, a random forest classification model on classifying classes of histology was trained and tested. A degree of accuracy higher than 90% was achieved. A second four-class model was trained to discriminate normal epithelium, malignant epithelium, normal stroma and cancer associate stroma. The classification accuracies are over 95% on the test data set.

Petersen et al. [51] presents the classification of several tissue components using a two-layer classification scheme based on Raman spectroscopic data. In the first layer, a multi-step random forest was applied, and it classifies the classes, connective tissue, muscle, erythrocytes, lymphocytes, crypts, carcinoma and lymph-follicle. In the second layer, linear discriminant analysis was used to reclassify carcinoma and lymph-follicle based on parameters from a spectral curve deconvolution algorithm. A pseduo-colour image was produced based upon the classification results. Further improvement of the image was achieved by overlaying Raman intensities of the C-H stretching vibrations with the Raman based random forest images.

Kuepper et al. [94] presented a two-layers random forest method which not only identifies tumour tissue with a sensitivity of 94% and a specificity of 100%, but also distinguishes cell differentiation and thereby tumour grading. The first

layer of the algorithm aims in the classification of cell types. The spectra that were classified as tumorous were further analysed in the second layer for cancer cell grading.

Detection and segmentation of histologic classes have been the major research focus in recent years. Most research on histologic class segmentation takes the following approach:

1. Build a spectra library containing different cell classes by the comparison of FTIR spectroscopy data and H&E stained image;
2. Divide the spectra library into a training set, and a testing set;
3. Train a supervised classification model for histological detection, using the training set;
4. Evaluate the classification model in the testing set;
5. Repeat step 3 to 4 to train several models, based on different supervised training methods;
6. Select the classification method has the best performance.

Cross-validation is used for a small spectra library to increase the data representativeness. The generalisability of much published research on this issue is problematic. There are two main reasons for that:

The repeatability of building a spectra library is low. In Fernandez's work, a library contains 171,000 pixels spanning all ten histologic classes was reported. Pilling used a spectra library consisting of 347,293 epithelium, 196,081 stroma,

8,151 blood, and 15,429 concretion spectra. Petersen gathered a spectra library contains following spectra: carcinoma (471), connective tissue (793), muscle (1538), erythrocytes (140), crypts (963), lymphocytes (593), lymph follicle (202), and background (1320). All the spectra in the library are selected by the trained histopathologist and spectroscopist. A high level of human work is required to build the spectra library, causing the low repeatability of the library itself and the low agreement between libraries.

Another fact that influences the generalisability significantly is the low representative of spectra. The size of cells can vary from about 5 to 30 $\mu$m, with the nucleus and organelles generally being 1 to 10 $\mu$m in size [52]. Typically commercial available FTIR spectroscopy has spatial specificity around 5.5 $\mu$m. Figure 9.1c is the H&E image of breast cancer tissue gridded by 5.5 $\mu m^2$ pixel size. In the figure, the measured spectra consist of an unpredictable mixture of spectral contributions from neighbouring cells. It is quite common to have mixture cell type in one pixel. Low spatial resolution combining with a manually sorting process, the spectral purity of the spectral library extracted from data is considerably low. The low accuracy spectral library can lead to low repeatability of the data analysis approaches.

An example of two classes missing to yield an entirely different one was published by Bhargava [48]. They found that pixels between epithelium and fibroblast-rich stroma are classified as mixed stroma in their cell type classification algorithm. The 'contamination' of the spectra caused by low spatial resolution was reported and discussed in detail by Fabian et al. [64]. To reduce the influence of low spatial resolution, different algorithms are developed to overcome

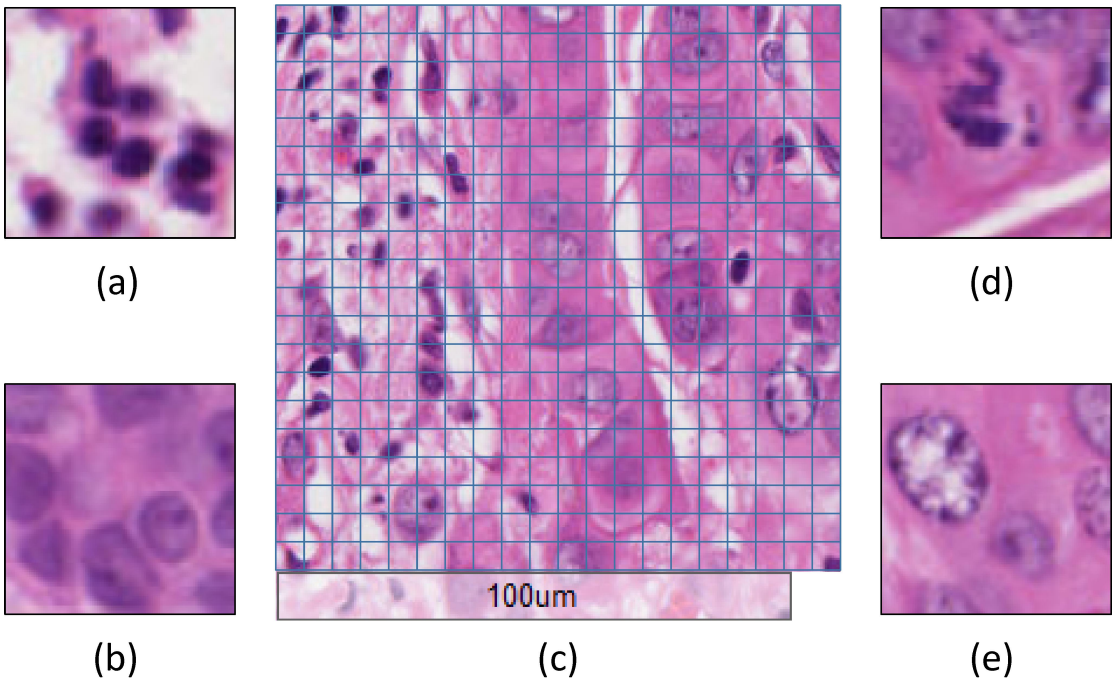



**Figure 9.1:** H&E image of different histologic cell classes and the dividing of one breast cancer tissue using a 5.5 $\mu m^2$ pixel size. (a) Lymphocyte nuclei (LN). (b) Epithelial nuclei (EN). (c) 20 by 20 dividing of one breast cancer tissue by pixel size 5.5 $\mu m^2$. Each square represents the spatial region that the spectrum corresponding to. (d) EN (Mitosis). (e) EN (Cancer). Figure produced based on sample 3, research identifier MI4502.

this problem. In Fernandez's work [73], probabilistic determination was used to reduce noise introduced by the low spectral representative in the training spectra set. In the random forest algorithm introduced by Pilling et al. [44], a probability of acceptance threshold of 0.6 was set to control the spectra purity problem. Petersen applied the mean and standard-deviation plot to confirm the consistency of each class and to observe the spectral difference between different classes [51]. Having an adjustable boundary classification method is the major similarity between those algorithms. However, as the representativeness of spectra various from library to library, the cut-off value of the classification

boundary varies. The reproducibility of those methods is not high.

### 9.2.1 Spectra library

The main method used to construct the library is by visually comparing the H&E image and the FTIR pseudo-colour image [44, 73]. To build the spectra library, first, different histological cell type regions were identified in the H&E stained tissue. Then, a pseudo-colour image of the tissue is generated using the FTIR imaging data. Cross-examination of the H&E image and the pseudo-colour image were performed to identify the corresponding spectra of a certain tissue region contains the same histologic cell. Those spectra were input into the library and marked as spectra of that histologic type. In this section, a preliminary spectra library was built, in the attempts of understanding the spectra selection processes.

The pseudo-colour image is generated based on the three-cluster Fuzzy C-mean clustering result of the first 30 PCs after applying PCA on the Amide I peak range (wavenumber range 1604 to 1697 $cm^{-1}$) of each spectrum. Mie-scattering correction, Savitzky-Golay smoothing and Amide II peak normalisation were applied in sequence as the pre-processing to the FTIR dataset.

In order to show a typical approaches I have undertaken, I have examined one of the tissue sample in the dataset and shown it below. The selecting processes of epithelial nuclei (EN) regions in the sample MI4502 is shown in Figure 9.2. As empty background is the region can be mostly correctly identified using PCA-FCM, an empty background region is selected as the calibration region.

By adjusting the magnification and the rotation angle of the two images, the calibration region is overlapped. The overlapping of the calibration region will automatically lead to the overlapping of other regions, as the H&E image and the pseudo-colour image are from two highly identical tissues. Using the identified empty background region, we can overlay the H&E image and the pseudo-colour image precisely (in Figure 9.2c) which helps to locate the corresponding spectra in the FTIR data set. The selected EN region is given in Figure 9.2d.

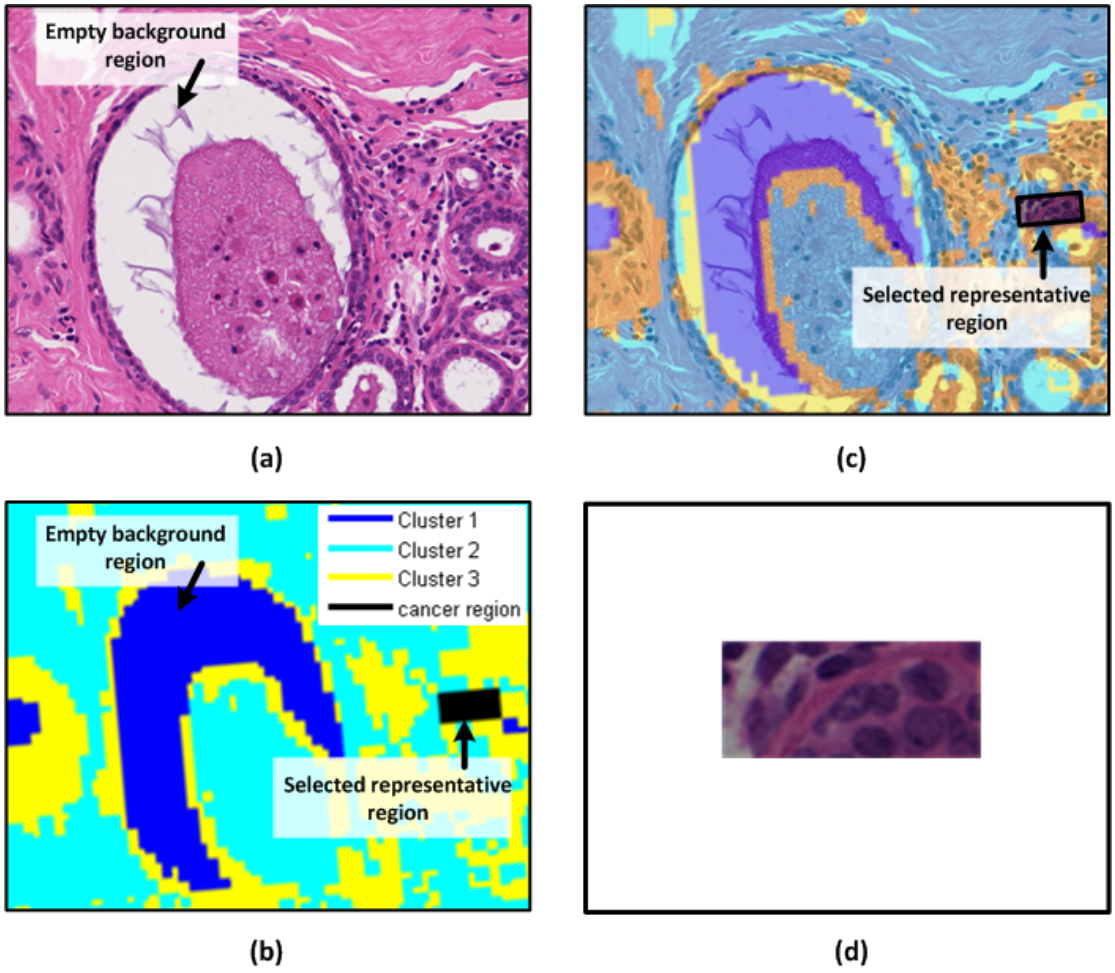



**Figure 9.2:** Selection of representative EN region (sample 3, research identifier MI4502). (a) H&E image; (b) pseudo-colour image on PCA-FCM; (c) overlapping of the H&E image and the pseudo-colour image; (d) H&E image of selected representative EN region.

The selecting process can be influenced by many things. First of all, the accuracy of the overlapping. Even though, PCA-FCM can detect empty background well, the shape of the calibration region for those two images (white region in Figure 9.2a and blue region in Figure 9.2b) still have a little difference, due to the resolution difference and the noise of IR spectroscopy data. Considering the unknown magnification and rotation angle of each image, it is difficult to precisely overlay those two regions. Second, the minor difference between the two tissue. The tissue for histological assessment and the tissue for FTIR spectroscopy data collection has high similarity. They are still two pieces of tissue which a certain level of difference can be found between them. Those differences make the correlation of histology classification with FTIR spectra difficult.

With much uncertainty involved in the constructing process, the spectra library is difficult to trust on a singular spectrum level. Compared with the size of cell, the 5.5 $\mu$m$^2$ pixel size is low, thus, the possibility to find single cancer pixel is low. Researches tend to select a homogeneous region of pathology instead of individual pixel [95]. We believe selecting a region containing the same histological type gives a higher confidence in the representative of the pixels in the library. In Figure 9.2d, even though the selected region is mainly covered by EN cells, it still contains stroma, background and blood cells. Those spectra collected from this region are corresponding to a mixture of histologic classes. It is difficult to tell the components of each individual spectrum, as it is difficult to locate an individual spectrum preciously. However, it is reasonable to suggest that those spectra as a group contains more EN spectra features than others.

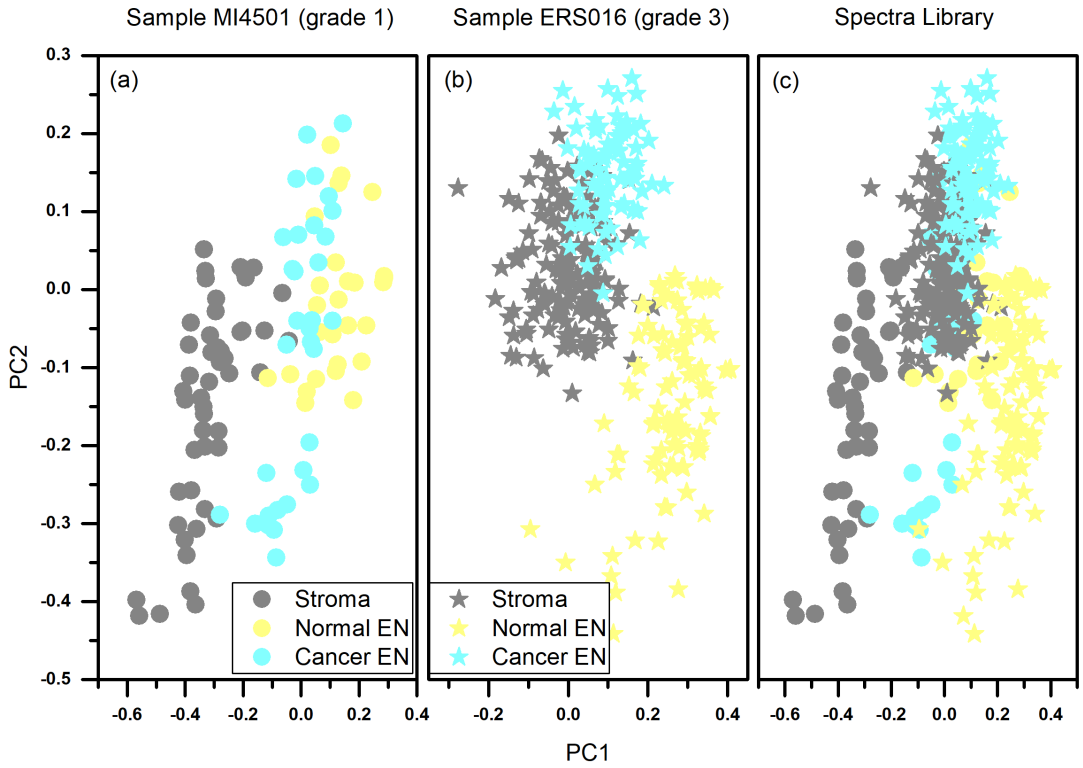



**Figure 9.3:** First two PCs scatter images based on the PCA analysis of Amide I peak in the spectra library. a) is the PCA result on grade one sample with each dot representing one spectrum. Gray, yellow and blue dots were used to represent spectra from stroma, normal and cancer EN respectively; b) is the same analysis result on grade three sample with each star representing one spectrum. Gray, yellow and blue stars were used to represent spectra from stroma, normal and cancer EN respectively; c) is the overlapping image of (a) and (b). Mie-scattering correction, Savitzky-Golay smoothing and Amide II peak normalisation were applied in sequence as the pre-processing to the FTIR spectroscopy data set.

Amide I peak range (wavenumber range 1604 to 1697 $cm^{-1}$) of the resultant spectra were subjected to PCA. Over 74% of the variance in the data set was described by the first two PCs. The scatter image is given in Figure 9.3. Figure 9.3a and Figure 9.3b is the first two PCs' scatter image of selected spectra for

sample MI4502 and MI4108 respectively. Clear clusters between stroma, normal and cancer EN can be observed. Stroma spectra focus on the bottom left. Cancer EN spectra located in the top middle while normal EN spectra all lie in the bottom right.

Figure 9.3c is an overlay image of (a) and (b). The scatter cluster between different histological classes still exists but less obvious. This demonstrates the spectra difference cause by histological reason can be masked by introducing the inter-sample variation. Figure 9.3c can be also taken as an evaluation of the spectra library. Based on the figure, in general, the spectra selected can represent the spectra characteristic of that histological type. However, due to the low accuracy of the selecting processes, the spectra library is unsuitable for individual spectrum comparison. To find spectra features of one histological cell type, it is better to use the whole spectra group other than an individual spectrum.

### 9.2.2 Detector edge, background subtraction

On the one hand, the current histological segmentation methods highly relying on the spectra library, which can be inaccurate; on the other hand, a certain level of segmentation is needed for bio-samples to increase the representativeness of the IR spectra. In this section, without relying on the spectra library, the empty/background pixels in the imaged tissue area, and the pixels collected by sensors lie on the edge of the focal plane arrays (FPA) detector were identified and cut off from the data cube. Those methods introduced in the following section are transparent, and can be easily related to chemical or physical reasons.

**Background cut**

The H&E image of the imaged tissue area (Figure 9.4a) clearly shows only the top-right corner is occupied by tissue. Pixels from the area containing no tissue structure are referred as empty pixels (containing no grading or pattern recognition information), which can provide optical distortion [96] and will conceal the real pattern or structure of the spectra data. Dividing those background pixels from the informative one and cutting them off is necessary.

Empty wavenumber range 1800-3000 $cm^{-1}$ in each spectrum was obtained, and the standard derivation of the first principle absorbance differentiation in this wavenumber range was generated. Two pixel columns (column one mainly contains empty pixels, and column two contains both empty and informative pixels) are picked, and the result of standard deviation is plotted in Figure 9.4b. Figure 9.4a is H&E image of the tissue showing the location of the two columns. For column two pixels, major differences between the background and informative pixels can be found. Comparing with empty pixels, informative pixels have lower standard derivation. A threshold of 0.0045 was used to divide empty from informative pixels.

The standard derivation difference observed in column 2 is due to the normalisation step in spectra pro-processing. Comparing with empty pixels, informative pixels have higher absorbance on Amide II peak. When normalising the spectrum, the absorbance on each wavenumber is divided by Amide II peak value. For informative pixels, the normalised absorbance value between 1800-3000 $cm^{-1}$ is lower as they are the result of the original value divided by a

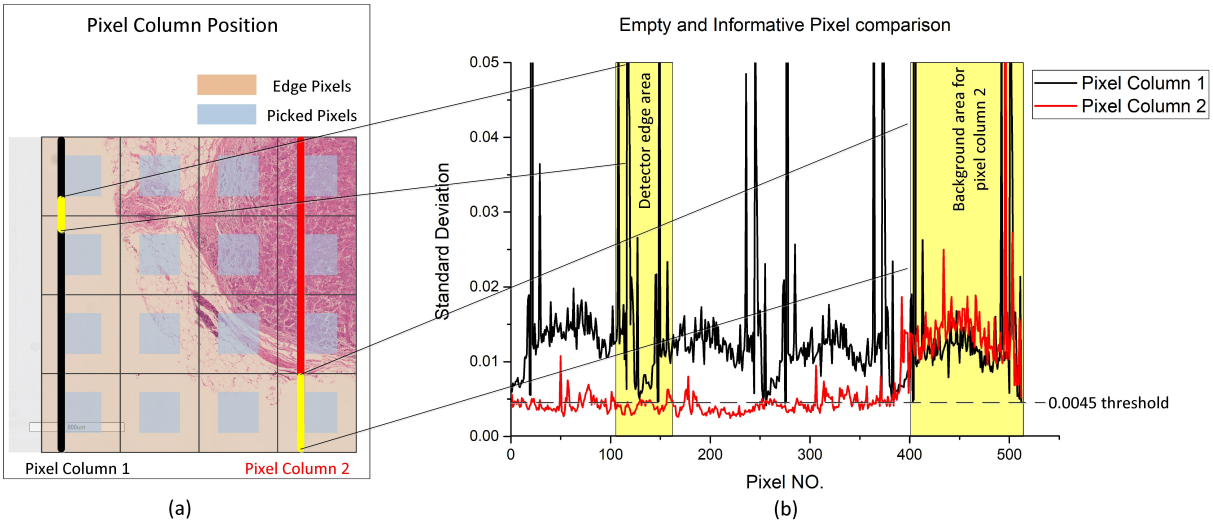



**Figure 9.4:** Standard deviation comparison of two pixel columns. (a) the H&E image of the tissue area. The positions of FPA detector during the FTIR data collecting steps are indicated by empty black squares ($4 \times 4$). The edge area and the central area of the FPA detector are marked as dark orange colour and light blue colour respectively. The positions of pixel column one and two are highlighted by black and red line respectively. (b) the standard deviation plot of pixel column one and two. Two interesting areas: the detector edge area in column one and the background area in column two, are highlighted. For pixels in column two, comparing with informative pixels, empty pixels have higher standard derivation value. For pixels in column one, a sharp dropping of the standard deviation value of pixels could be found at those pixels, which lie on the edge of the focal plane arrays detector (detector edge area). Tissue sample used in this figure is from grade three sample MI 4108.

relatively high Amide II peak value, which will further lead to lower standard derivation, compared with empty pixels.

Figure 9.5 is the comparison of H&E image and the pseudo-colour image of grade one sample MI4502. Using the standard derivation 0.0045 as classification criterion, pixels with lower standard derivation are classified as informative

pixels and coloured as yellow, meanwhile, pixels with higher or equal to 0.0045 standard derivation are classified as empty pixels and coloured as blue. The pseudo-colour shows that the method can provide accurate background pixel identification. The H&E and pseudo-colour image comparison of other samples agree with this conclusion.

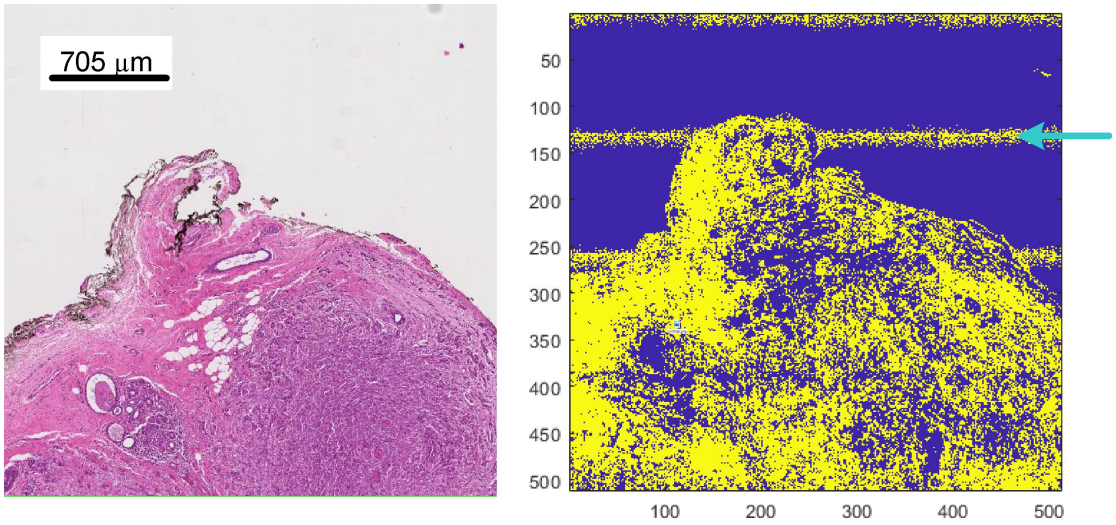



**Figure 9.5:** Sample MI4502 (grade one) background cut H&E image and pseudo-colour image comparison. On the left, is the H&E image of the imaged region. On the right, is the pseudo-colour image generated based on the background cut result, with yellow represents tissue pixels and blue represents empty pixels. The miss identified region caused by the detector edge problem is highlighted by the light blue arrow.

An interesting patter that can be observed in the pseudo-colour image of sample MI4502 (Figure 9.5) is that the background segmentation method has high mis-classification rate on pixels collected by the edge of the detector. A location-based edge spectra segmentation is introduced.

**Edge cut**

Care must be taken to avoid pixels that lie on the edge of the detector as these may artificially inflate the error [7]. There are many reasons that can cause the changing of data quality across the FPA detector. A possible explanation for variation data quality across the FPA detector can be the relatively short life-cycle of the detector. FPA detector was initially developed to be used as a missile guidance sensor, in which, it is designed to be used once only. The lifetime of majority commercialised FPA is still short. It is possible that, the FTIR spectroscopy machine used for collecting the image data has an FPA that is reaching its time limits, which results in the low performance on the detector edge.

The real reason for the low S/N ratio spectra collected by the edge sensors is still unknown. One thing for sure is that the variation is not due to any bio-chemical composition difference, as the misclassification exists for pixel column one, which are mainly empty pixels. Pixels that lie on or close to the detector edge were cut off before further analysis is applied.

Evidence of changing of data quality of edge spectra can be seen in both Figure 9.4 on page 210 and Figure 9.5 on page 211. In Figure 9.4b, sharp changing of standard deviation value could be found at those pixels lie on the edge of the focal plane arrays detector (highlighted region of the line plot). In Figure 9.5, the empty pixels lie on the detector edge are misidentified as informative pixels due to the low S/N ratio.

An edge cut algorithm was developed based on the spatial location of each

spectrum. For each 128 $\times$ 128 sub data cube, twelve pixels rows/columns from the four edges of the FPA detector square were cut off. The remaining spectra form a 104 $\times$ 104 data cube. In Figure 9.4a on page 210, each black square represents one sub data cube created by the FPA. The region in dark orange colour is recognised as the edge pixel region. The selected pixels for further analysis are highlighted in blue. The background segmentation pseudo-colour image and standard derivation study of other samples confirmed the algorithm could effectively identify those low S/N ratio spectra lie on the detector edge.

Based on the tissue structure of the imaged area, the number of spectra remained after background, and edge cut varies from sample to sample. Table 9.2 is the summary of the number of spectra been selected for further analysis. The ratio of the number of spectra selected to the total number of spectra collected (512 $\times$ 512) is provided.

All over the 30 samples, on average, 86250 spectra were selected. Those selected spectra were further analysed to provide histological grading information of each tissue. Grade three sample ERS016 has the lowest select-ratio. The further analysis indicated that the tissue sample tends to be an outlier in most analysis methods.

**Table 9.2:** 30 sample dataset spectra segmentation summary (RI=Research Identifier; No=Pixel Number; R=Ratio). For most samples, large number of spectra were selected for further statistic analysis.

| RI | No | R | RI | No | R | RI | No | R |
|---|---|---|---|---|---|---|---|---|
| ERS 091 | 46191 | 0.18 | ERS 027 | 130664 | 0.5 | ERS 009 | 64954 | 0.25 |
| MI 4197 | 119121 | 0.45 | MI 4258 | 53447 | 0.2 | ERS 013 | 31382 | 0.12 |
| MI 4502 | 63507 | 0.24 | MI 4373 | 123107 | 0.47 | ERS 016 | 18760 | 0.07 |
| MI 4531 | 155234 | 0.59 | MI 4569 | 105523 | 0.4 | ERS 018 | 107384 | 0.41 |
| MI 4572 | 99405 | 0.38 | MI 4758 | 104327 | 0.4 | ERS 024 | 97177 | 0.37 |
| MI 4690 | 116153 | 0.44 | ERS 069 | 82766 | 0.32 | ERS 076 | 52397 | 0.2 |
| MI 5061 | 107985 | 0.41 | ERS075 | 118850 | 0.45 | ERS 102 | 69247 | 0.26 |
| MI 5103 | 104734 | 0.4 | ERS 109 | 102339 | 0.39 | MI 4108 | 57730 | 0.22 |
| MI 5267 | 78505 | 0.3 | ERS 111 | 42105 | 0.16 | MI 4117 | 101033 | 0.39 |
| ERS 014 | 41290 | 0.16 | MI 4107 | 91558 | 0.35 | MI 4121 | 100632 | 0.38 |
| Average pixel number: 86250; Average ratio:0.33 | | | | | | | | |

CHAPTER 10

# Dimension reduction

One of the difficulties of FTIR spectroscopy data analysis is the processing of the high dimension data. FTIR spectroscopy measures the IR absorbance on a range of wavenumbers. For different step size, the number of variables in each spectrum varies from dozens to hundreds. A typical mid-IR spectroscopy spectrum contains IR absorbance on wavenumber range 600-4000 $cm^{-1}$. Using step size 2 $cm^{-1}$, each spectrum contains 1700 variables. Each variable will be considered as one dimension in the data analysis steps. Not all variables have value for cancer diagnosis or prognosis. Inputting all the variables into the analysis algorithm will increase not only the computational expense but also the noise level. Identifying and selecting a set of principle variables for IR spectroscopy data is essential and critical. This process is called dimension reduction in data mining.

Knowledge-based FTIR spectroscopy data dimension reduction is possible. For instance, in Chapter 5 virgin polypropylene with carbon fibre section, standard polypropylene and surfactant IR spectra were used to identify 'polypropylene

peak' and 'surfactant peak' to reduce the dimension of the spectra. Breast cancer FTIR spectroscopy data knowledge-based dimension reduction, however, is difficult as the prior knowledge required about the biochemical composition of breast cancer cell is still under investigation. Pruning of biological FTIR spectroscopy data, in which the complex compositions are unknown and changing from case/person to case/person, requires an in-depth understanding of the data structure and high level of data analysis skills. Machine error and noise of the data are making this process a more delicate job. Patterns can be easily found, however, it is hard to find patterns that have logical meaning.

Teh et al. [83] introduced random forest to near-infrared Raman spectroscopy data analysis on the larynx sample for the first time. The random forests method was introduced to develop effective diagnostic algorithms for classification of Raman spectra of different laryngeal tissues. The result, indicated that NIR Raman spectroscopy in conjunction with powerful random forests algorithms have a great potential for the non-invasive, in vivo diagnosis and detection of malignant tumours in the larynx at the molecular level.

In a recent study, a two layers random forest workflow was created by Kuepper et al. [97] for precise colorectal cancer tissue classification based on quantum cascade laser-based infrared microscopy. They analysed 110 patients with UICC stage II and III colorectal cancer, showing 95% sensitivity and 100% specificity compared to the golden standard in routine clinical diagnosis. The tissue level cancer diagnosis was achieved by setting a 2% threshold of denoting tumour cells percentage.

A paper by Pounder et al. [7] was the first study trying to exam the potential to

combine the spectral and spatial information for breast cancer diagnosis. Each pixel was first labelled as stroma or epithelium using spectral recognition at the single pixel level; subsequently, epithelium pixels are labelled as cancer or normal by spatial polling based upon epithelium content and distribution. Results indicated that the protocol is highly accurate under a variety of conditions.

A widely accepted procedure of multi-wavenumber analysis for biological FTIR spectroscopy data is as following:

1. Generate spectra metrics by selecting the wavenumbers that have the most variation between histologic classes;
2. Apply classification on each spectrum to group them into histologic classes;
3. Evaluate the classification result using the true classification from pathology;
4. Repeat step 1 to 3 for several times;
5. Select the best performance metric based on the classification result.

IR spectra shape can easily be influenced by various sample preparation and data collecting, pre-processing methods. The spectra changes caused by cancer relevant biochemistry change is small. It can be easily masked by those noises. The noise deduction pre-processing steps have limited ability in increasing the signal-to-noise level. Highly human involved analysing together with the large dynamic range of IR spectra, made the selected best performed metric various from study to study.

The developments in computational power allow the spectra pathology without metric selection. Instead of selecting the metric manually, the whole spectra

were input to algorithms such as random forest [44, 51] or deep learning [98] to train a self-learning algorithm to perform tissue pathology, whom is specialised for dealing with high dimension data. The best performing model was selected to conduct tissue segmentation on the test sample set. Using a high dimension data process algorithm, this approach successfully reduced the level of human involvement, meanwhile increases the objectiveness. However, both random forest and deep learning have limited mathematics support. This makes physically or biologically interpreting of the FTIR spectroscopy data histology result impossible. The potential of understanding cancer biochemistry development through FTIR spectroscopy is lost.

In this chapter, the difficulty of conducting dimension reduction on breast cancer IR spectroscopy data is clearly stated by a spectrum study on a knowledge-based approach analysis of PP and PE IR spectra. Evidence that combining multiple wavenumber can improve the S/N ratio is provided right afterwards to further support the needs of using multiple wavenumber absorbance for IR analysis. The benefit of applying probability density distribution analysis on breast cancer tissue IR spectroscopy image data is detail discussed. Different dimension reduction methods were explored. Traditional multivariate statistical analysis PCA combining curve fitting and multivariate statistical analysis on breast cancer FTIR spectroscopy data was tried in Section 10.5. Section 10.6 is the application of the spectrum symmetric analyses algorithm to reduce the dimension of the data. In Section 10.7, based upon a cancer development model, standard score analysis was introduced as a dimension reduction method.

## 10.1 Spectrum changes caused by one single methyl group

This part aims to provide a simple example, that illustrates how IR spectrum changes when the chemical information changes in the material.

Compared with breast cancer tissue sample, the plastic sample has a simple structure and relatively uncomplicated spectrum. A large and growing body of literature has investigated the infrared spectra of high polymers making the knowledge-based interpretation of the infrared spectrum possible. The polyethylene (PE) and polypropylene (PP) IR spectra analysis are simple examples of IR spectra analysis. From PE to PP, only one change of the molecule (H to $CH_3$) is made. The resulting spectrum shows rather complex peak changes (rising, falling, appearing, disappearing and shifting) on multiple wavenumber.

Examination of PE and PP using FTIR spectroscopy has been extensively studied [99–102]. The assignments for most of the spectrum bands have already been satisfactorily made [101]. The molecular structure of PE can be considered to be an approximation of an infinite chain of $CH_2$ groups. In comparison to a PE polymer, a PP chain has a $CH_3$ on every other carbon group. Figure 10.1 displays the structure of molecules of PE and PP.


**(a)** Polyethylene **(b)** Polypropylene

**Figure 10.1:** Molecular structure of polyethylene and polypropylene. In comparison to a PE polymer, a PP chain has a $CH_3$ on every other carbon group.

Figure 10.2 is the infrared spectrum of virgin PP. The spectrum is the average of 36864 (192 $\times$ 192) spectra collected from PP film using FTIR imaging technique. Based on the study result reported by Krimm et al. [101], infrared spectrum peaks of PE and their corresponding assignment of chemical bonds vibration were marked on the spectrum. Compared with PE, PP contains more $CH_3$ bonds, which will lead to higher absorption in the corresponding $CH_3$ IR spectrum peaks. In Figure 10.2, the rising of $CH_3$ composition is causing peak disappearing (1353, 1470, 2853 and 2925 $cm^{-1}$), appearing (1359 $cm^{-1}$) and peaks maxima position frequency shifting (1456, 2874 and 2959 $cm^{-1}$).

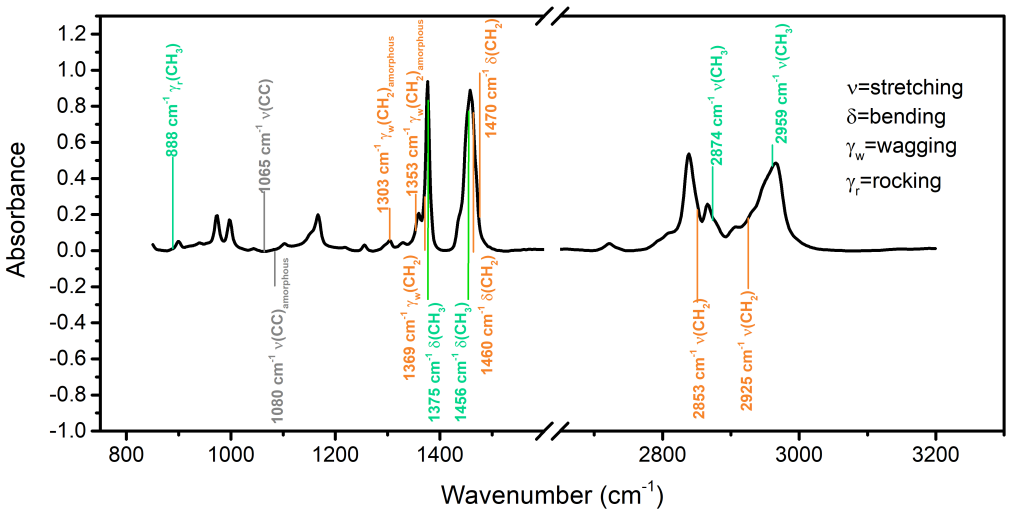



**Figure 10.2:** Infrared spectrum of PP. Infrared spectrum peak and their assignment were marked based on a spectrum study on PE. Comparing with PE spectrum, peak appearing, disappearing and shifting can be observed due to the rising of $CH_3$ composition.

For breast cancer cell, as the development of cancer, the chemical bonds changing inside of the cell are much more complex and robust. Different changes in chemical bonds give different spectrum changes. It is difficult to select mul-

tiple wavenumber on a knowledge-based approach as the biochemical changes caused by cancer are still under study. Knowing a certain type of protein is highly relevant to cancer, the IR spectroscopy study of that protein is most likely lacking. This is one of the reasons that dimension deduction of biomedical IR spectroscopy data is difficult.

Furthermore, the number of one type of change in the focused IR beam region is much less compared with the PP, PE example in which every other molecule is having the same H to $CH_3$ change. The dropping of the number of one type change in breast cancer sample will lead to the spectra changes of that molecular change getting smaller. IR spectrum is the overlaying of all the IR absorbance relating to all chemical bonds containing in the specimen. The spectra changes between cancer and normal cell are the superposition of hundreds or even thousands of small spectra changes. At most wavenumber the absorbance is changed in a non-naked-eye-observable level. This makes the dimension reduction on a data-driven approach difficult.

The PCA, symmetric and zscore analysis introduced in this chapter are all trying to increase the information representative meanwhile decrease the dimension of each spectrum without wavenumber selecting step.

## 10.2 S/N ratio and multi-wavenumber

In FTIR spectroscopy data, the signal desired is the IR absorbance corresponding to chemical structures. The higher the composition, the bigger the IR absorbance. Meanwhile, the machine and the background noise level remain con-

stant across all wavenumbers. The signal-to-noise (S/N) level varies between wavenumber to wavenumber. Selecting those wavenumber has a higher S/N ratio as input, can improve the S/N ratio level of that spectrum. A standard procedure for IR spectrum analysis is truncating the spectrum to only remaining the wavenumber has a high S/N ratio.

In applying FTIR spectroscopy for cancer prognosis or diagnosis, the signal desired is further narrowed down to cancer reverent IR absorption. It has been widely accepted, that by combining multiple wavenumber absorbances, cancer-relevant variation percentage (S/N ratio) can be increased [98]. Increasing number of studies reported the clinical application of IR spectroscopy based on multiple wavenumbers. Machine learning or manually comparison is often used, to find the best combination of wavenumber based on the performance.

Corresponding to a certain chemical structure, IR absorbance spectrum is normally a Lorentzian shape peak in a wavenumber range. Within this range, the S/N ratio changes continuously. Suppose, there is a 'cancer IR peak', because the continuously changing of S/N ratio, the peak position and its sounding wavenumber should be selected. The state of the art wavenumber selection results are mainly individual, non-continuity values. It is difficult to interpret those result with biological or medical reasons.

The molecular fingerprint region (600-1800 $cm^{-1}$) is considered to have a high S/N ratio for IR spectroscopy cancer development features. In the following sections of this chapter, Amide I peak (wavenumber range 1604 to 1697 $cm^{-1}$), the single most widely used amide modes of protein secondary structure [103], is selected as the input. Different algorithms were tried to abstract cancer de-

velopment information based on this region.

## 10.3 Probability density distribution on FTIR spectroscopy data analysis

To have an idea of the 'middle' or centre of a data set, several statistical concepts could be used, with mean, median and mode the most popular. Mean is the average value, giving the same weight to all the data points. However, it can be easily influenced by the extremely large or small outliers. Median is the middle value of the data set. It lacks the ability to give a general picture of all the data points. For instance: [1, 2, 3] and [-100, 2, 100] have the same median but huge difference in mean and standard derivation. Mode is finding the most common value. It is not applicable for samples have continuous numeric sample space.

For a sample that has complex components, for instance, human tissue, none of the three are enough to give a clear description of the centre distribution of the FTIR data. This section is a detailed explanation of why probability density distribution was adopted as the main method to describe FTIR imaging data in this thesis.

### 10.3.1 Variation in tissue FTIR spectroscopy data

Tissue contains up to 10,000 different proteins and all kinds of circulating molecular species such as sugar, lipids, peptides and metabolites [104]. The concentrations of different components have large dynamic ranges across the tissue sample.

For tissue FTIR spectroscopy data, the variation of one spectrum contains can be generally grouped into the following groups: variation from the biochemical components, variation from environment/machine noise and variation from sample preparation. The latter two can be removed or reduced by the pre-processing of the FTIR spectroscopy data. Variations of the biochemical components are directly related to cancer diagnosis or prognosis.

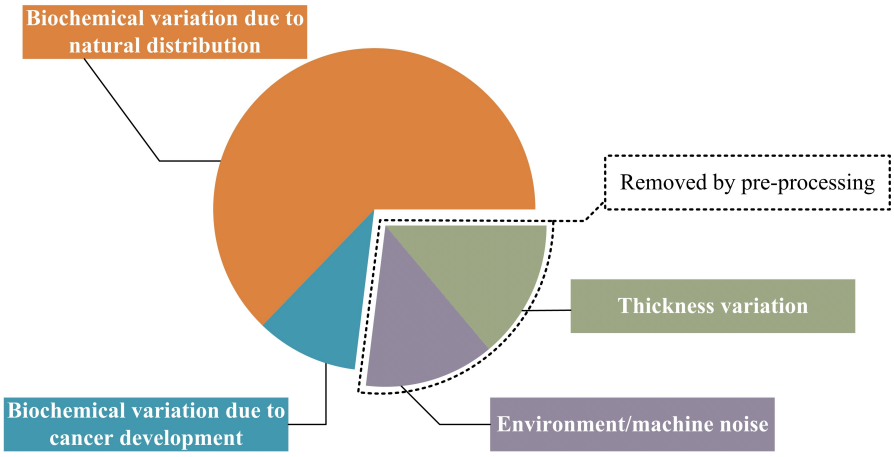



**Figure 10.3:** Schematic of the cancer tissue spectra information. The information containing in one spectrum normally comes from the following four: biochemical variation due to natural distribution, biochemical variation due to cancer development, Environment/machine noise and thickness variation.

Due to the components variation between cells as well the mixture of more than one cells in one pixel (Figure 9.1 on page 203), the biochemical components are different from pixel to pixel. Cancer development can also cause the biochemical changes of tissue. After the pre-processing step, the spectral difference observed between spectra is corresponding to the biochemical difference, which can be further split into variation from natural uneven distribution and variation caused by cancer development. Figure 10.3 is the schematic of the

information containing in the cancer tissue spectra. Difference can be easily observed between spectra. However, it is difficult to tell the difference observed is due to natural uneven distribution or cancer development.

### 10.3.2 The power of statistics

Due to the complexity of the tissue components, singular IR spectrum quantitatively study on cancer development is difficult. In this thesis, statistical analysis was introduced to extract cancer information on a population level. Density distribution is introduced as a stronger statistic tool to represent the trends in IR absorptions.

The reason for using statistical analysis for tissue FTIR data is as following:

1. It is difficult to control the biochemistry components on a singular spectrum level. Human traits vary across a wide range; however, in most cases, taking cancer cells in one tissue as a population, if we graph the number of cells, which are in a certain level of cancer developments, we'll get a distinctive chart known as a normal curve.

2. The IR absorption changes caused by cancer are real but quite small. They are so subtle that we can be easily masked by the other variations. This is, one of the technical difficulties to conduct cancer analysis based on individual spectrum comparison. Statistically significant, however, is a numerically reliable way to claim a difference between two spectra population. The distribution between cancer and normal tissue can be highly overlapping, as major components of the two tissue are still the same, but

difference can be observed by a statistical tool, as they are, statistically belonging to different populations.

3. Spatial variation is needed to achieve cancer diagnosis or prognosis. The general distribution of different types of cells within the tissue is essential information for evaluating the cancer stage of the patient.

Taking each pixel as one individual, pixels from one tissue as the group of individuals, the difference between groups representing the difference between tissues. If the major difference between two pieces of tissue is the cancer level, the variation observed on the population level can be directly associated with cancer developments.

Details of statistical analysis of FTIR tissue image data are given in Figure 10.4. Starting from the 3-D matrix resulting from FTIR spectroscopy, dimension reduction is applied to reduce the dimension of one pixel into one value, which contains the biochemical information of that pixel. That single value is called the feature value of that pixel. After obtaining the feature value, the outlier cut was applied before generate the density distribution. $Quartile1$ - (1.5 $\times$ $interquartilerange$) and $quartile3$ + (1.5 $\times$ $interquartilerange$) were used as the boundary to identify outliers. Bin the feature value over one sample which breaks a high amount of continuous feature value into a series of density values (estimated based on the normal kernel function). Each of the bin-result-value represents the density value integrated over a small range. After binning, first-term Gaussian fit was applied to the density value to obtain the probability density distribution. The fitting of density distribution is detailed explained in

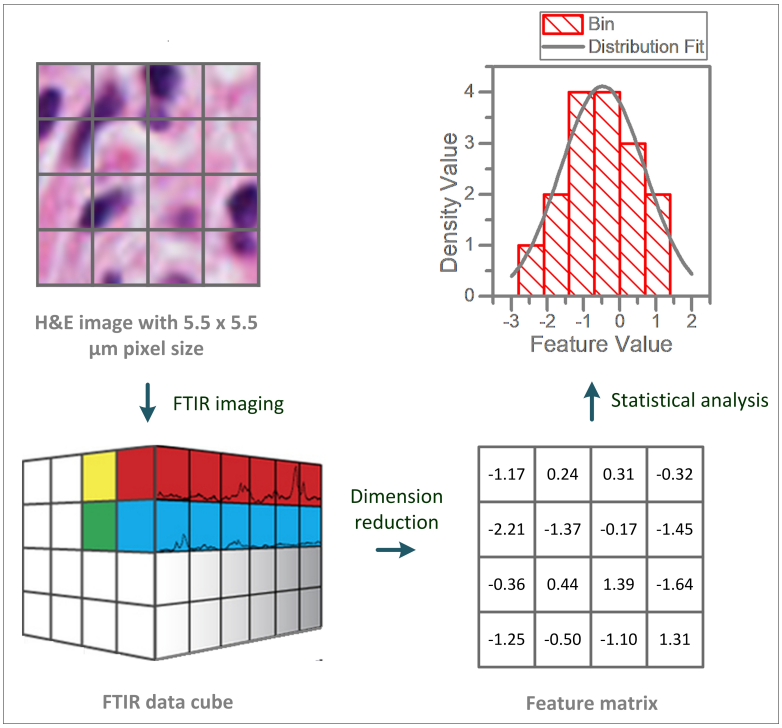



**Figure 10.4:** Statistical analysis of FTIR tissue image data. A spectra matrix was created by the data collecting step. Each spectrum in the matrix contains the absorbance variation of the materials in that pixel. Dimension reduction was applied on the spectrum level to reduce the z axis to one single feature value. The density distribution of that value across the tissue was then calculated and plotted. Density distribution function was then obtained by regression.

next Section 10.4. A step-by-step instruction of obtaining the density distribution by binning is given in Section 10.5.1.

For non-cancer tissue, as lone as the population size is large enough, the feature value across a tissue should be following a normal distribution. The existence of cancer cells in the tissue can twist the normal distribution into a different shape. By this statistical analysis approach, the biochemical variation across one piece of tissue can be summarised by a distribution function. The biochemical variation due to natural uneven distribution should be a normal distribu-

tion. Between tissue samples, this normal distribution should be the same as the physical reason behind them is natural uneven distribution. Any difference we observed on the density distribution, is due to the biochemical variation on a population level. Cancer developments can cause the changing of biochemical components in a population of cells. In other words, the cancer level of those tissues can be revealed by the density distribution of the feature value.

## 10.4 Regression on FTIR spectroscopy data analysis

Regression is a statistical method that is popularly used in data mining. Relationship between variables is estimated by a mathematics equation which is used to describe the changing of dependent variables based on the independent variable. The benefit of applying regression is as following:

**MAKE USE OF PRIOR KNOWLEDGE.** There are huge amounts of functions can be used to estimate the relationship between variables, such as: Fourier, Gaussian and polynomial models. To apply a certain kind of function requires the prior knowledge of the relationship. For example, for IR absorbance spectra, after many studies and comparison, Lorentzian, Gaussian and Voigt functions are normally used to fit spectral line shape [105]. The prior knowledge on the shape of the relationship can be introduced to the data when applying a certain regression function.

**REDUCE NOISE.** Regression is describing the major form of relationship based on pairs of dependent and independent variables. Random noise or machine error could be excluded from the data set if predictions of the dependent variable

are made. This is a standard and rather developed approach in data mining. Petersen et al. [51] used a fifth order polynomial to fit each spectrum to remove the residual spectral baseline of Raman spectroscopy data.

### 10.4.1 Regression on feature value density distribution

In this thesis, for all the FTIR imaging data set, regression is used to link the binning result with density distribution function to enable the comparison of feature value on a population level. As shown in Figure 10.4 on page 227, after abstracting and binning feature value, regression is applied to the binning results to describe the distribution mathematically.

Considering the feature value as a random variable, we denote it by symbol capitalised litter $X$. A real feature value data point is $x$. Study the probability for the event $X < x$. This probability is a function of $x$ and is called the distribution function of $X$:

$$F(x) = P(X < x) \tag{10.4.1}$$

When this distribution function $F(x)$ is continuous and differentiable, the function's first derivative can often provide special interesting information:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = F^{'}(x) \tag{10.4.2}$$

is called the probability density (function) of $X$. It is a measure of the probability of the event $(x \leq X < x + \mathrm{d}x)$. From 10.4.1 and 10.4.2 it follows that

$$P(X < a) = F(a) = \int_{-\infty}^{a} f(x)\mathrm{d}x \tag{10.4.3}$$

$$P(a \leq X < b) = F(b) - F(a) = \int_{a}^{b} f(x)\mathrm{d}x \tag{10.4.4}$$

and in particular

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{10.4.5}$$

Normal distribution (Gaussian distribution) is a common continuous distribution. It is often used in biology and social science to represent real value variables whose distributions are not known.

The density distribution caused by natural uneven distribution tends to be a normal distribution. Supposing we have two large enough samples from the same population, the density distribution should be identical normal distributions. As discussed before, after pre-processing, the IR absorbance contains two kinds of variation only: variation from natural uneven distribution and variation caused by cancer development. Any difference can be observed from the density distribution is corresponding to biochemical changes in cancer development.

First-term Gaussian fit (Normal distribution fit) is applied to the distribution to obtain the distribution function. The function is given:

$$y = a \cdot e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2} \tag{10.4.6}$$

where, $\mu$ is the mean or expectation, $\sigma$ is the standard derivation of the distribution.

In the following chapter of this thesis, for FTIR imaging data, the parameters of the first-term Gaussian regression function on probability density distribution are used to represent the sample.

### 10.4.2 Regression on spectral line shape

The second usage of regression in this thesis is in Section 10.6, symmetry analysis of Amide I peak. Taking wavenumber as an independent variable while the absorbance of that wavenumber as a dependent variable, a vibrational spectrum can be estimated using a mathematics equation. Regression is used to describe the shape-changing of Amide I peak. Symmetric level of the peak was abstracted based on the regression results.

## 10.5 Principal components analysis

Principal component analysis (PCA) is one of the most widely used data mining approaches for dimensional reduction. It reveals the most prominent variation patterns in data by transferring the data into a different PCs space using an optimal eigenvector matrix [106].

PCA can estimate the most common patterns across the features [107]. In Section 9.2.1, applying PCA to selected spectra, differences between histologic classes were revealed. If most absorbance variations of the spectra data are changing with respect to cancer status, the first principal component will be

highly correlated with cancer status.

In this section, taking each spectrum as one observation, PCA was applied to Amide I peak (wavenumber range 1604 to 1697 $cm^{-1}$) of selected spectra after background and edge cut (section 9.2.2). The first PCs value was selected and feed into the probability distribution function fitting. Parameter of the probability distribution function was used to compare the first PC value distribution between samples.

The preprocessing methods for data sets in this section are: RMieS-EMSC $\rightarrow$ spectra truncation $\rightarrow$ Savitzky-Golay smoothing $\rightarrow$ Amide II peak normalisation. The effects of the normalisation step were evaluated by applying the same PCA transfer to both normalised (data set identifier in coding: ZCorr_normal) and non-normalised (data set identifier in coding: ZCorr_smooth) data set.

### 10.5.1 Single PCA

After cutting the empty and edge pixels off, all over the 30 samples, on average, 86250 spectra were selected per tissue sample. Taking each spectrum as one observation, combining all the 30 samples' selected spectra (observation) into one matrix, PCA was applied.

The transferred observations were re-grouped into 30 clusters based upon the sample they are from, before further statistic analyses were applied. The result data set is 30 clusters of observation which are transferred into the new PCs space from the original spectra space. Figure 10.5 is the schematic illustrating the single PCA process.

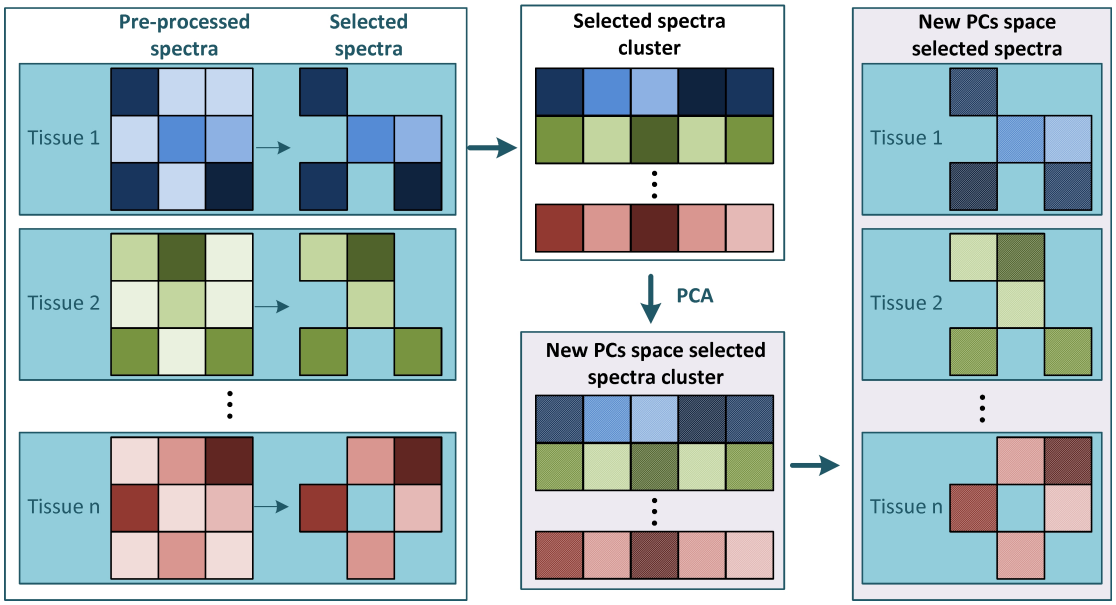



**Figure 10.5:** Single PCA schematic. On normalised (coding identifier: ZCorr_normal) data set, the selected spectra from all 30 samples were joined into one 2-D matrix, with each row of the matrix representing one spectrum (observation). PCA was applied to the matrix to transfer it into the new PCs space. Observation in the new PCs space was then re-grouped based on the tissue sample it is from.

After PCA, density distribution analysis was applied to each of the 30 samples. The first PC, which takes most of the total variation, was selected to represent each spectrum/observation. For each of the 30 tissue samples, following steps were used to obtain the density distribution of the first PC value:

1. Apply a standard outlier cut. *Quartile one* - (1.5 x *inter quartile range*) and *quartile three* + (1.5 x *inter quartile range*) were used as the boundary to cut off outliers.

2. Generate a row vector XI, which contains 100 linearly equally spaced points between the minimum and maximum value of the variable. (MATLAB function: linspace)

3. Based on a normal kernel function, using a window parameter (bandwidth) that is optical (Equation 10.5.1), the probability density is estimated from the sample in the variable. Vector XI is used to specify where the density estimate is to be evaluated. The estimated result is given in the form of cumulative probability values. (MATLAB function: ksdensity)

The optimal bandwidth, in step 3, is given [108]:

$$bandwidth = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}} \tag{10.5.1}$$

where, $\sigma$ is the standard derivation of the sample, $n$ is the number of points.

In Section 10.5, 10.6 and 10.7, after abstracting new feature values, the density distribution of that feature value, within each tissue sample, was obtained using those steps described above.

First-term Gaussian fit was applied to the distribution to obtain the distribution function. Parameter of the function was scatter-plotted to compare the difference between samples. The result and the discussion of the result are given in Section 10.5.3. To test normalisation's influence on FTIR spectroscopy data, the same single PCA analysis was applied to non-normalised data set (data set identifier in coding: ZCorr_smooth). The density distributions of the first PCs value of each spectrum were obtained. The Gaussian fit results were given in Section 10.5.3.

### 10.5.2 Multi-layer PCA

For cancer tissue FTIR spectroscopy data, the variation can be divided into two types: inter-grade and intra-grade variation. The inter-grade variation is the IR spectra difference caused by cancer developments. The intra-grade variation (difference of samples with the same cancer grade) is the IR spectra difference caused by sample variation, sample preparation or data collection process, which is not cancer relevant.

For cancer prognosis, in the ideal situation, PCA transfers the original spectra data into a PCs space where the inter-grades difference is maximised while the intra-grades difference is minimised. In scatter results, we want the difference between grades large; meanwhile, the difference between samples in the same grade small.

Applying PCA directly will transfer the data into s PCs space that the major difference is maximised. It is difficult to identify the source of this major difference, therefore, making the claim the major difference abstracted is cancer relevant (inter-grade difference) is difficult. A double-layer PCA method is proposed in this section, to provide more control over PCA transformation.

Based on the eigenvector matrix, PCA transfers the data into another PCs space. Applying the same eigenvector matrix can transfer other observations into the same PCs space. Normally, the eigenvector matrix is obtained from the sample space directly. Abstracting the eigenvector matrix from a sample space that the inter-grade difference is the major variation can give the eigenvector matrix ability to pick out grading information and maximise it.

Pixels in the same sample went through the same treatments. The IR spectroscopy data is collected in the same environment and machine setting. Both the sample preparation and data collection variation is low in-between those observations. For the 30 tissue sample FTIR spectroscopy data, within each of the tissue sample data cube, the cancer development information is included, while the intra-grade variation is minimised.

In a double-layer PCA process, the first layer is abstracting the eigenvector matrix using observations from one tissue sample only. In the second layer, apply the same eigenvector matrix to perform PCA to other tissue sample data.

By this double-layer PCA approach, PCA was forced on finding the spectral difference of cancer development than the difference between samples.

Double layer PCA was conducted using the following steps:

1. Layer one: Apply PCA on the observation from one sample. Tissue sample 1 (ERS091) is used in this step. (Using other tissue samples in these steps has been tested. The influence of the choice of tissue samples in this step to the result is little.)

2. Layer two: Using the same eigenvector matrix as layer one, transfer the FTIR spectroscopy data of other samples into the same PCs space.

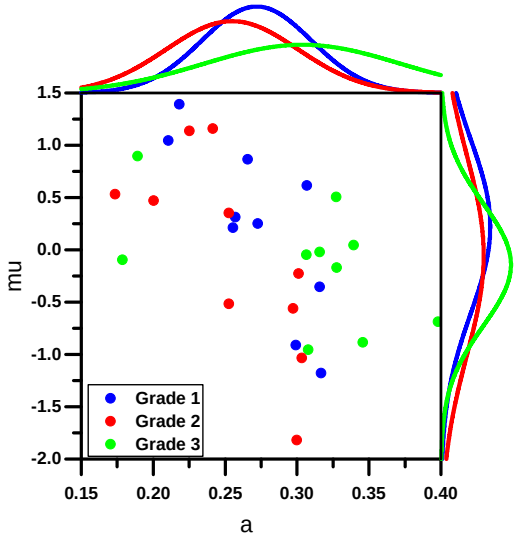
Abstracting eigenvector matrix from sample 1 (grade 1), double-layer PCA was conducted. Density distribution analysis was applied to each of the 30 samples. The first PC, which takes most of the total variation, was selected to represent each spectrum/observation. Based on first-term Gaussian curve fit, the probability density functions of each sample were obtained. The $a$ and $\mu$ parameter

scattering image is given in Section 10.5.3. To test normalisation's influence on FTIR spectroscopy data, the same double layer PCA analysis was applied to non-normalised data set (data set identifier in coding: ZCorr_smooth). The density distributions of the first PCs value of each spectrum were obtained. The Gaussian fit results and the detailed discussion of the results were given in Section 10.5.3.

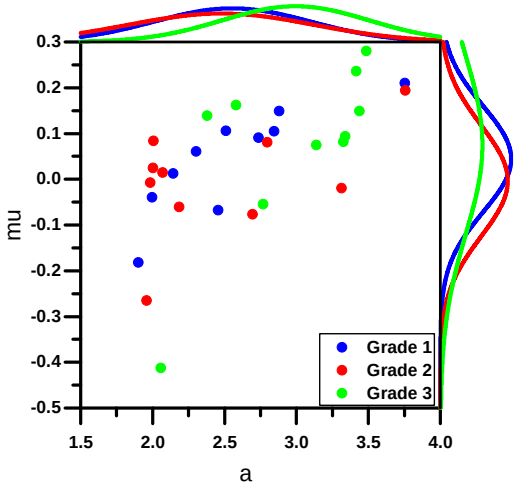
### 10.5.3 Result and discussion

After obtaining the feature value (the first PC of the spectrum), density distribution analysis was applied to each of the 30 samples as described in Section 10.5.1 on page 232. First-term Gaussian fit (Normal distribution fit, as described in Equation 10.4.6 on page 230) was applied to the distribution to obtain the distribution function. The parameters of the first-term Gaussian regression function on probability density distribution were used to represent the sample. The $a$ and $\mu$ parameters scattering image of single and double-layer PCA on normal and smooth data set are given in Figure 10.6.

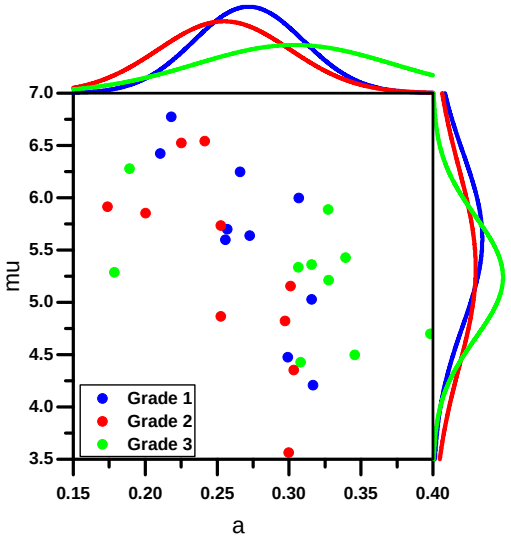
Except for the $\mu$ parameter range (y axis), scatter image 10.6a and 10.6c looks almost identical. This indicates both single and double layer PCA identified the same variation as the major variation in the non-normalised data set. As double-layer PCA is designed to abstract eigenvector matrix in a low inter-sample variation space, this major variation identified by both PCA method may due to the natural uneven distribution of cellular contents. This result highlighted the importance of applying normalisation to bio-sample FTIR spectroscopy data to reduce the influence of natural uneven distribution.



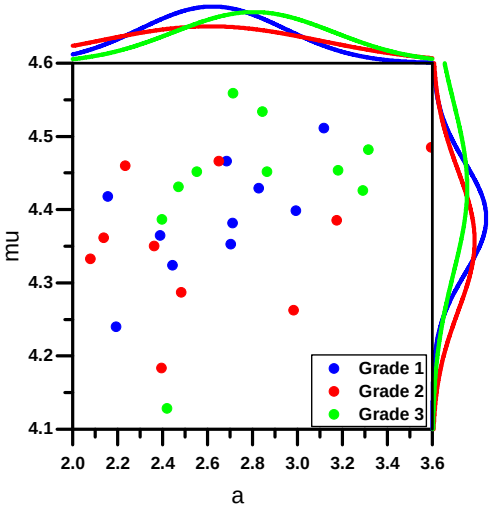

**(a)** Single PCA on smoothed data set.




**(b)** Single PCA on normalised data set.




**(c)** Double PCA on smoothed data set.




**(d)** Double PCA on normalised data set.

**Figure 10.6:** Density distribution function parameter scatter, with blue dots representing tissue samples from grade one, red dots representing tissue samples from grade two and green dots representing tissue samples from grade three. The scatter plot $a$ and $\mu$ with marginal distribution curves. The marginal distribution curves take samples from the same grade as a sub-population. The distribution curve reveals the difference of distribution pattern on each of the dimensions.

The conclusion is in good agreement with the pre-processing methods reported in the literature, and it can be double confirmed by comparing single PCA performance on non-normalised (10.6a) and normalised data set (10.6b). Compared with non-normalised data set, single PCA on normalised data can provide relatively better dividing between cancer grades.

Figure 10.6b and 10.6d provided the performance comparison of single and double-layer PCA on normalised data set. In both results, certain differentiation between different cancer grades can be observed. Grade three samples tend to be in the top-right corner while grade one samples are in the bottom-left. Grade two samples are mixed in-between of grade one and three samples.

Compared with single PCA, double-layer PCA can transfer the data into the PCs space that less inter-sample difference can be observed. However, it is difficult to make the conclusion which one performs better, as both single and double-layer PCA are not providing good enough separation of samples from different grades. This may due to PCA methods are transferring the data into the PCs space that intra-grade variations are big and inter-grade variations are small, causing the later one is being masked. The low performance of PCA methods may also be caused by the normalisation step. As discussed in Section 6.1, applying peak value normalisation can introduce a high level of noise to the data.

As the PCA transferred data lost its original physical meaning, it is difficult to study the actual cause of the low performance and improves it. New dimension reduction methods are introduced in the later sections of this chapter, attempting to provide transparent, interpretable feature extraction methods for FTIR

spectra data.

## 10.6 Symmetric analysis

De Meutter et al. [49] used arrays of infrared detectors to obtain high resolution images of protein microarrays. 100 $\mu$m protein spots each containing about 100 pg protein were deposited to form high density regular arrays. Spectra of the 16 proteins are presented in Figure 10.7. It can be seen that the spectra difference between proteins can be large enough to be observed (carbonic anhydrase and Hemoglobin), or too small to apply visual interpretation (Lysozyme and Myoglobin).

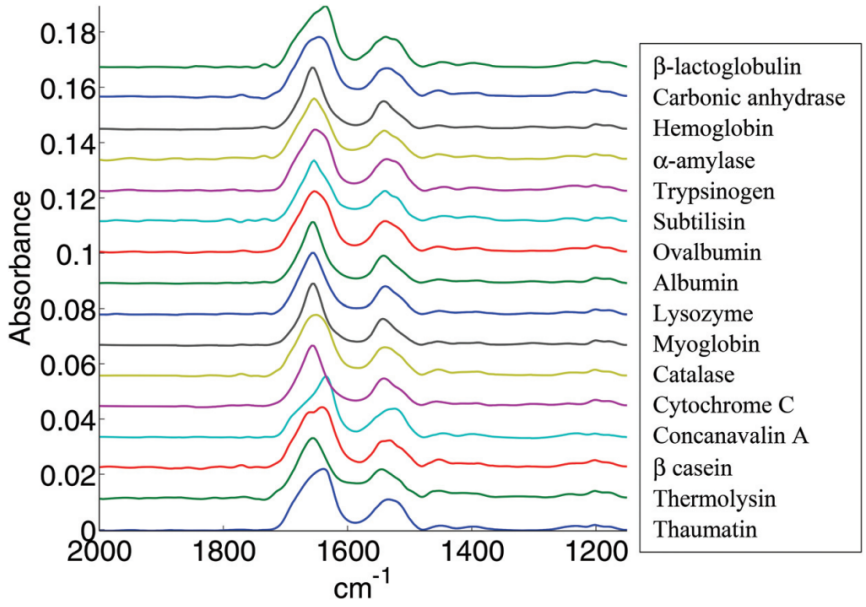



**Figure 10.7:** Mean spectra of the 16 proteins obtained. Spectra have been offset for the sake of readability. Reproduced from [49].

The development of cancer can lead to biochemical composition changing in cells, which can cause the appearing, disappearing, falling or rising of peaks in

IR absorbance. The composition changes are so small that the spectral changes caused by cancer development are difficult to apply visual interpretation directly. A symmetric-analysis method is proposed in this section to reveal the composition changing based on the changing of the band shape.

In this section, different mathematics equations were tried to describe Amide I peak's rising and falling edge. The best-fit equation was found, and the parameter of the equation was used to describe the symmetric level of the peak. Then, for the selected pixels in each sample, the distribution of the symmetric level was obtained. Probability density distribution function was used to show the general symmetric level of that sample.

The preprocessing methods for data sets in this section are: RMieS-EMSC $\rightarrow$ spectra truncation $\rightarrow$ Savitzky-Golay smoothing. Focusing on the shape-changing of Amide I peak, normalisation step does not change the symmetric level of the peak, which makes the peak symmetric analysis results same between normalised and non-normalised data set. The symmetric analysis results are directly comparable with PCA results, as the algorithm is based on the same preprocessing methods and pixel selecting process.

### 10.6.1 Why Symmetric analysis

Protein can fold into complex three-dimension structures, which consist of a variety of domains containing polypeptide segments folded into different types of secondary structures, which have their individual IR absorption. The observed amide I bands containing many overlapping component bands that rep-

resent different structural elements such as $\alpha$-helices, $\beta$-sheets, turns, and non-ordered or irregular structures [103]. Model calculation on Amide I bands of globular protein has demonstrated that other $\alpha$-helices, $\beta$-sheets, protein secondary structure spectral contributions spread over wide wavenumber range [109]. Increasing number of experimental observations indicated the difficulty of identifying individual composition and de-composition amide I band [110].

The method currently used in extracting the information is based on the 'principle of pattern recognition'. Using machine learning or other data analysis approaches, taking each wavenumber absorbance as individual variables, patterns of the variables were selected. This approach does not require that individual bands be assigned to different type of secondary structure. It, however, requires a high level of mathematical manipulation of the data which makes it difficult to interpret the results. Thus, it does not address the fundamental problem of the lack of clearance in the relationship between amide bands and the type of secondary structure.

By applying peak symmetric analysis, focusing on the shape-changing on a series of wavenumber instead of discrete wavenumber absorbance value, unobservable changes of the spectra are revealed. Together with the statistic distribution method, the major trends of the spectra shape changing are obtained. The benefits of applying symmetric analysis can be summarised as following:

- A continuous wavenumber range, other than discrete wavenumber is used to in the multi-wavenumber algorithm, without the need for band decomposition. The reason for using multi-wavenumber and the arguing of con-

tinuous multi-wavenumber is better than discrete is presented in Section 10.2.

- The symmetric analysis algorithm is based on the smooth data set (data set identifier in coding: ZCorr_smooth), in which the absorbance value is used as the input variables. It makes the analysis result more interpretable. This can be used to improve the understanding of cancer developments.

As far as we understand, this is the first report of applying peak symmetric analysis on FTIR spectroscopy data. This provides an optional asthmatic method for spectra analysis. A thorough application of this method can form a useful supplement to the existing methods in crystallising biological interest of cancer development.

### 10.6.2 Amide I peak rising and falling edge fitting

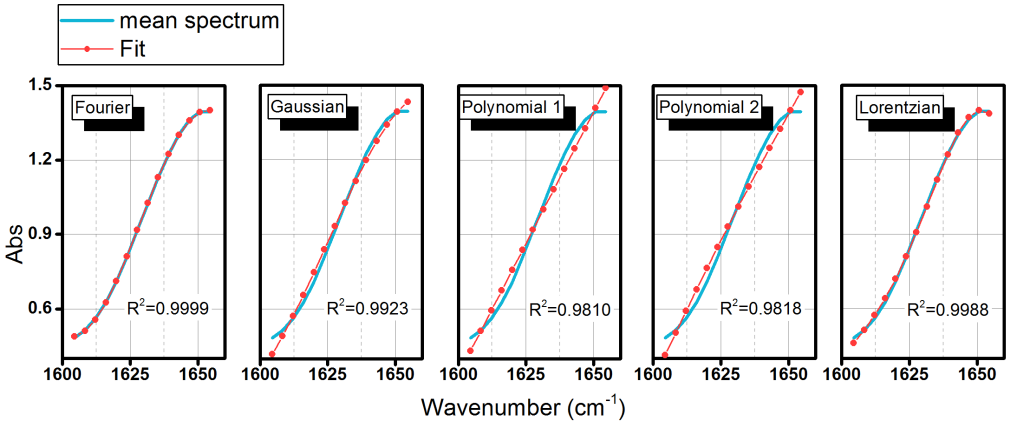
Early researchers used multiple terms of functions to fit IR spectra to reveal the components of examined material [17, 111]. With spectacular accuracy rate, curve fitting is limited by its low objectiveness. Series subjective decisions (terms of the function, peak position of each term) need to be made before a result can be achieved.

The shape-changing of Amide I peak during the development of cancer has been reported [112–114]. Even though those multivariate analyses could give similar results as curve fittings, the further application of the results is hindered by its low linkage with biochemistry knowledge. It is difficult to interpret multivariate analysis results with biological changes.

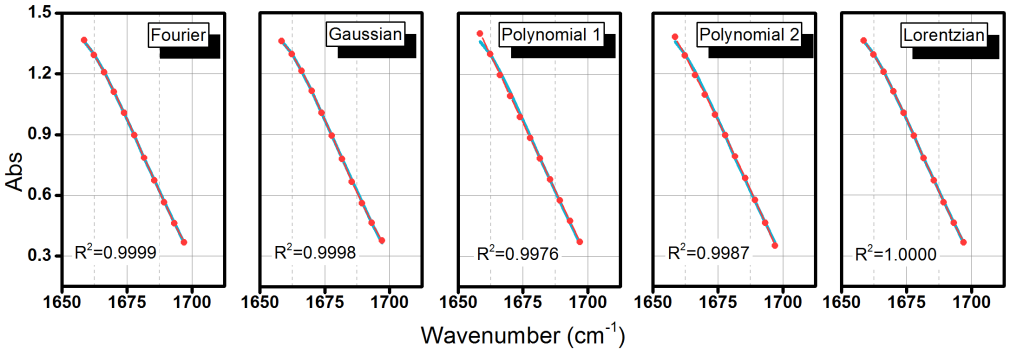
In the symmetric analysis methods we propose, Amide I peak (wavenumber range 1604 to 1697 $cm^{-1}$) was selected as the input wavenumber range to analysis the symmetric level of the peak. Curve fitting methods were applied to the rising and falling edge of Amide I peak respectively to get two individual mathematics equations. The peak antisymmetric level was obtained from the parameter of those two individual mathematics equations. One term mathematics equation curve fitting was chosen as the result is interpretable using biochemical knowledge.

To find the best regression function for Amide I peak, one-degree polynomial, two-degree polynomial, first-term Gaussian, first-term Fourier and Lorentzian functions were used to fit the rising and falling peak edge. Taking the wavenumber as an independent variable, the mean absorbance of all the selected cancer EN (Section 9.2.1) on that wavenumber as a dependent variable, different curve fitting methods were performed. The results from the fitting can be seen in Figure 10.8 and Table 10.1.

Both Lorentzian and Fourier function fits well on Amide I peak. For rising edge, Fourier fit performs better than Lorentzian with $R^2 = 0.9999$. For falling edge, Lorentzian fit can provide $R^2 = 1$ result while Fourier fit's $R^2$ is 0.9999. The difference of the fitting performance between the rising and falling edges could be the direct result of spectra components difference in the wavenumber range. Compared with the falling edge which has a shape similar to a single component spectra line, rising edge shape is closer to the shape of Fourier function, which indicates the rising edge has more robust spectral components. This finding is in line with those of previous studies [112–116].




**(a)**




**(b)**

**Figure 10.8:** Amide I peak fit. (a). rising edge fit; (b). falling edge fit. In both, from left to right, first-term Fourier, first-term Gaussian, one-degree polynomial, two-degree polynomial and Lorentzian functions were used to fit the mean spectrum rising edge (a), and falling edge (b). Blue line is the cancer EN mean spectrum of the spectra library from Section 9.2.1. The red line with dots is the fitting results.

**Table 10.1:** Cancer EN mean spectrum fit results with the best fitted highlighted in red.

| | Fit | | Goodness of fit | |
|---|---|---|---|---|
| | Fit type | Fit result | SSE | R-square |
| Rising edge | Polynomial (degree1) | $y = 0.0212x - 33.51$ | 0.02937 | 0.9810 |
| | Polynomial (degree 2) | $y = -0.00005x^2 + 0.1689x - 153.8$ | 0.02805 | 0.9818 |
| | Gaussian (term 1) | $y = 1.463e^{-(\frac{x-1662}{51.54})^2}$ | 0.01196 | 0.9923 |
| | Lorentzian | $y = 0.0335 + 1.366\frac{63.54^2}{4(x-1651)^2+63.54^2}$ | 0.00147 | 0.9988 |
| | Fourier (term 1) | $y = 0.9441 - 0.2042\cos(0.0620x) + 0.4082\sin(0.0620x)$ | 0.00017 | 0.9999 |
| Falling edge | Polynomial (degree1) | $y = -0.0267x + 45.75$ | 0.00280 | 0.9976 |
| | Polynomial (degree 2) | $y = -0.00008x^2 + 0.2536x - 189.4$ | 0.00147 | 0.9987 |
| | Gaussian (term 1) | $y = 1.417e^{-(\frac{x-1650}{40.73})^2}$ | 0.00029 | 0.9998 |
| | Lorentzian | $y = -0.922 + 2.358\frac{52.71^2}{4(x-1649)^2+52.71^2}$ | 0.00003 | 1.0000 |
| | Fourier (term 1) | $y = 0.7971 + 0.6927\cos(0.0420x) - 0.0695\sin(0.0420x)$ | 0.00012 | 0.9999 |

Lorentzian fit relies highly on the initial starting point. The fit performance heavily influenced by the subjective decision made by the user. Even though the fitting performances of Lorentzian and Fourier function are similar, first-term Fourier fit was selected as the best-fit function for further symmetric analysis.

### 10.6.3 Symmetry

The IR spectrum changes in the following two ways: the increasing or decreasing of chemical bonds will cause the corresponding fingerprint peak rising or dropping; the appearing or disappearing of chemical bonds can lead to the corresponding appearing or disappearing of fingerprint peaks. Section 10.1 Spectrum changes caused by one single methyl group, provides a detailed example of the rather complex spectra changes corresponding to one methyl group on the polymer chain. For breast cancer cell, as the development of cancer, the composition changing inside of the cell are much more complex and robust. This is the fundamental reason that band de-composition is unsuitable for biology FTIR spectroscopy data analysis.

A peak in one spectrum (e.g. Amide I peak) is normally the integral of multiple chemical bonds fingerprint peaks within that region. As discussed above, if we cannot assign the absorbance difference to each of the chemical species, we should take the peak as a whole in-dividable piece. The increasing, decreasing, appearing or disappearing of chemical bonds fingerprint peak can cause the changing of symmetric level of the spectrum peak. An example of the symmetrical level changes with the peak composition, is given by Coleman and Zarian

[117]. As shown in Figure 10.9, the peak symmetrical level changes when the poly($\epsilon$-caprolactone) (PCL) and poly (vinylchloride) (PVC) blend has a different percentage composition.

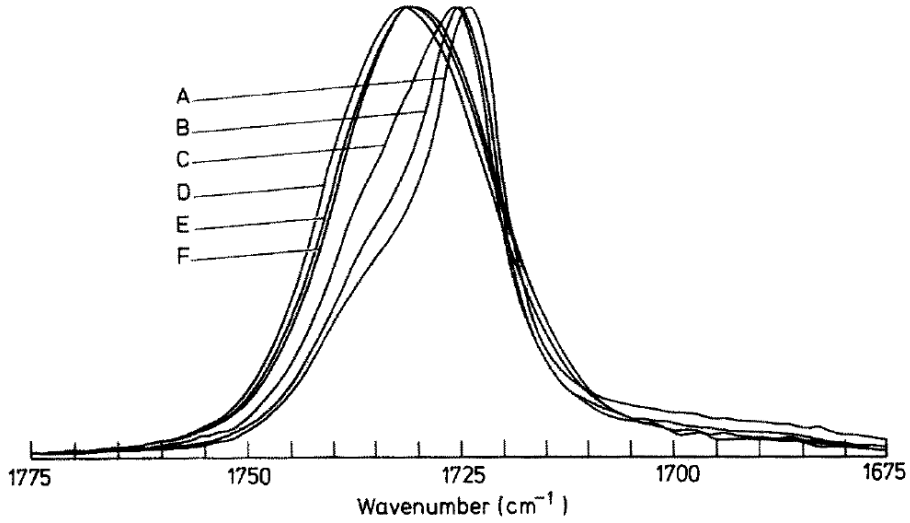



**Figure 10.9:** FTIR spectra of PVC-PCL blends recorded at room temperature in the range 1675-1775 $cm^{-1}$. (A) Pure PCL, (B) 1:1, (C) 2:1, (D) 3:1, (E) 5:1, (F) 10:1 molar PVC:PCI, respectively. Reproduced from [117].

To better illustrate the relationship between spectrum peak symmetric and biochemical components changes, a highly simplified and hypothetical model is given in Figure 10.10 on page 250. The model is a de-composition of spectra peak in the wavenumber rang 1600-1700 $cm^{-1}$. It assumes a Lorentzian shape for the original components [17], and to maintain the simplicity only two bands with the same width were considered. For chemical bond 1, the position of maximum is 1630 $cm^{-1}$, and the full width at half maximum is 40 $cm^{-1}$. For chemical bond 2, the position of maximum is 1655 $cm^{-1}$, and the full width at half maximum is 40 $cm^{-1}$.

Assuming the developments of cancer makes chemical bond two reminds the same while chemical bond one increases, the spectrum and the decomposition of normal and cancer cells are given in Figure 10.10a and Figure 10.10b respectively. One-term Fourier fit was used to fit the spectrum rising and falling edge. The results show that in normal cell, the rising edge has a fitting parameter $\omega$ equals to 0.046 while the falling edge's value is 0.052. The gap between rising and falling edge is 0.006. In cancer cell, the rising edge's $\omega$ increases to 0.048 while the falling edge drops to 0.049. The gap between rising and falling edge is 0.001. The model shows that the $w_{rising} - w_{falling}$ parameter can be used to describe the symmetric level of a peak, thus show the components changes within the tissue sample.

Using single oscillating function, namely Sines and Cosines, one-term Fourier function was found to fit Amide I peak best. The detailed investigation can be seen in Section 10.6.2 The mathematical function of one-term Fourier model is:

$$f(x) = a_0 + a_1 cos(wx) + b_1 sin(wx) \tag{10.6.1}$$

in which $x$ represents wavenumber value, $f(x)$ is the IR absorption level in the wavenumber of $x$. Four parameters are used: $a_0$, $a_1$, $b_1$ and $w$. Parameter $w$ defines the frequency of the oscillating function which is highly relevant with the shape of the peak. By subtracting the $w$ between the rising and falling edge fit function, we can get the symmetric level of the peak.

Following is a step by step instruction of getting the Amide I peak symmetric level of each spectrum:

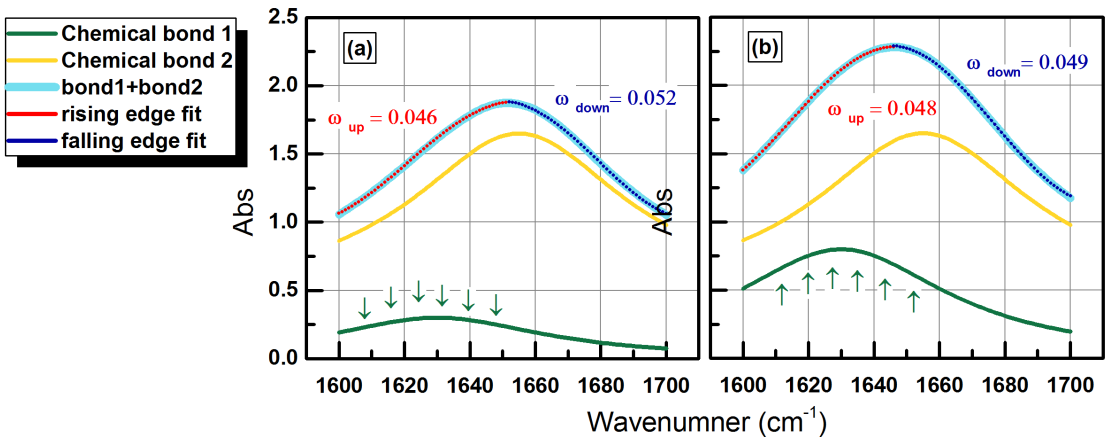



**Figure 10.10:** Relationship between chemical bond fingerprint peak, spectrum peak and symmetric level. (a) Normal cell spectrum decomposition. Chemical bond one and two are given in green and yellow line respectively. The integrations of these two peaks are given in blue line. Using one-term Fourier fit, the rising edge and falling edge's fit parameter $\omega$ is given. (b) Cancer cell spectrum decomposition. Compared with normal cell, chemical bond one increase while chemical bond two remains the same. The changes in the spectrum components can be revealed by the $\omega$ parameter of Fourier fit.

1. Select Amide I peak (wavenumber range 1604 to 1697 $cm^{-1}$). Divide the peak into rising and falling edge from peak position.

2. One-term Fourier fit on the rising and falling edge of Amide I peak.

3. Subtract the $w$ parameter of the rising and falling fit function.

Fourier fit regression can summarise the shape of the Amide I peak rising or falling edge into one single value. Symmetric analysis compares the value of the rising and falling edge from the same spectrum in which it contains the same variation from noise and thickness. By subtracting the two $\omega$ values, we removed the thickness and noise variation, and subtracted biochemical inform-

ation.

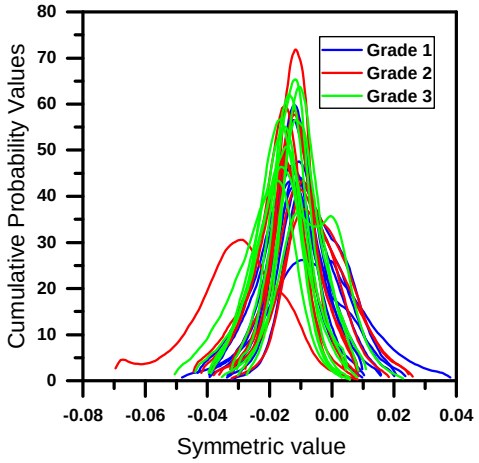
### 10.6.4 Symmetric value density distribution

Applying symmetry analysis to a spectrum, the whole spectrum can be represented by $w_{rising} - w_{falling}$, which is called the feature value of that spectrum. After the feature extraction process, the 3-D data cube is transferred into a 2-D feature matrix, with each spectrum represented by its own feature value. Bin the antisymmetric value ($w$ subtraction) within each sample. A Gaussian shape, density distribution function can be obtained based on the result of the bin. This process is illustrated in detail in Section 10.3, Figure 10.4 on page 227.

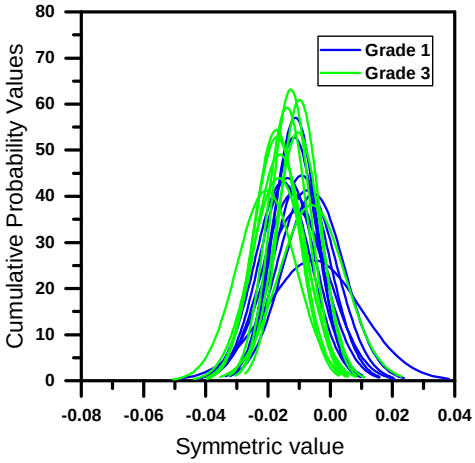
The bin results, the density distribution (Gaussian fitted bin results) and the density distribution function scattering are given in Figure 10.11. The bin plot of grade one samples tends to be flatter and have higher peak position compared with grade three bin plots. Comparing with pixels in these grade three samples, spectra from grade one samples tend to have symmetric value closer to zero, which means the rising and falling edge of Amide I peak has closer $w$. Along with the cancer developments, the cellulite component changing causes the reduction of the symmetric level of the Amide I peak.

The shape of the density distribution is related to the variation level. The sharper the shape the lower the variation is. The density distribution of the symmetric value tends to be shaper in grade three comparing with other grades, which means a high uniform level of symmetric in higher-grade cancer.

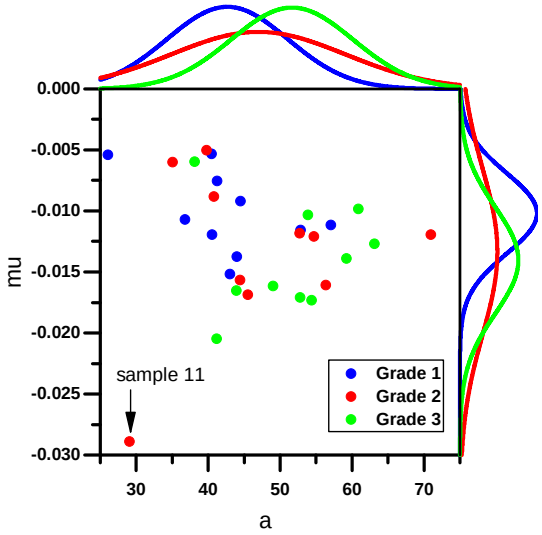
Feeding the density distribution into Gaussian fit to further reduce the dimen-



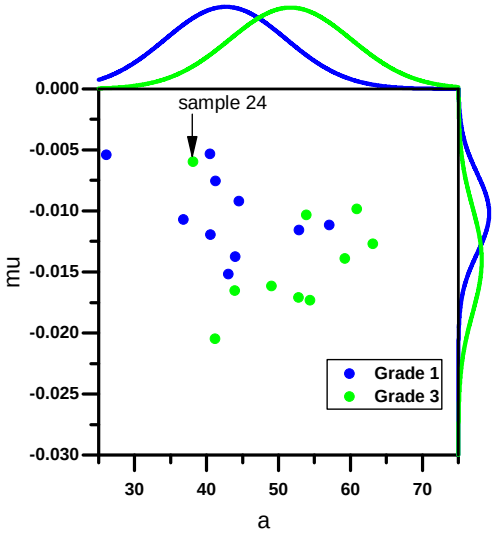

**(a)** Bin plot.




**(b)** Gaussian fitted bin plot.




**(c)** Density distribution function parameter scatter.




**(d)** Density distribution function parameter scatter with grade 1 and 3 samples.

**Figure 10.11:** Symmetric analysis results with colour code: blue (grade one), red (grade two) and green (grade three). (c) and (d) is the scatter plot $a$ and $\mu$ with marginal distribution curves. The marginal distribution curves take samples from the same grade as a sub-population. The distribution curve reveals the difference of distribution pattern on each of the dimensions. Samples from grade one and three are divided with grade two tissue distributed in-between.

sion of each sample, the density distribution function of each sample is obtained. Parameter $a$ and $\mu$ of each Gaussian fit are chosen as the two variables representing each sample, the scatter plot of the function parameters is given in Figure 10.11c. Figure 10.11d is the same scatter plot with only grade one and three samples. Grade one and grade three samples are in two different clusters while grade two samples are mixed in between.

Outliers were identified in both scatter plot. There are many possible reasons for the outliers: the low data quality (S/N ratio) or the miss-classification by NPI.

Symmetric analysis of Amide I peak provides a rather sensitive approach of detecting spectral changes due to protein secondary structure. Any structure changes of protein secondary structure, both those have or have not been fully explored by protein study, can be taken into count when cancer prognosis is made based on FTIR spectroscopy spectra. It can be useful as a data-driven knowledge discovery tool to highlight cancer relevant IR spectra change, which further improves cancer protein study. One thing that worth noting, other components of the cell, e.g. amino acid side-chain [118], also provide spectral contribution into Amide I peak region.

## 10.7 Standard score

In the previous section, by applying symmetric analysis on Amide I peak, cancer tissue samples from grade one and three can be divided. This is satisfactory, as it is, as far as we know, the first breast cancer study based on FTIR spec-

troscopy reporting the dividing of different pathology grades. However, comparing with the golden standard NPI, this grading result is less exciting as the NPI has a high accuracy rate in terms of prognosis on grade one and grade three. Identifying grade two breast cancer from other cancer grades is relatively difficult. The misclassification of grade one is one of the aspects that NPI performance can be improved.

In this section, a rather common normalisation/standardizing method, standard score/zscore, was applied to attempt to divide grade two breast cancer samples. To gain a better understanding of the relationship between cancer developments and molecular changes, a simple cancer development model was built, based on which, zscore was proposed as a dimension reduction method. Standard score was used to reveal the IR spectra changes between cancer grades. Clear cluster of grade two with other grades is observed from the results.

The preprocessing methods for data sets in this section are: RMieS-EMSC $\rightarrow$ spectra truncation $\rightarrow$ Savitzky-Golay smoothing. The zscore analysis results are directly comparable with PCA results (on smooth data set) and symmetric analysis results, as the algorithm is based on the same preprocessing methods and pixel selecting process.

### 10.7.1 Introduction to standard score (zscore)

In statistics, the standard score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. Observed values above the mean have positive

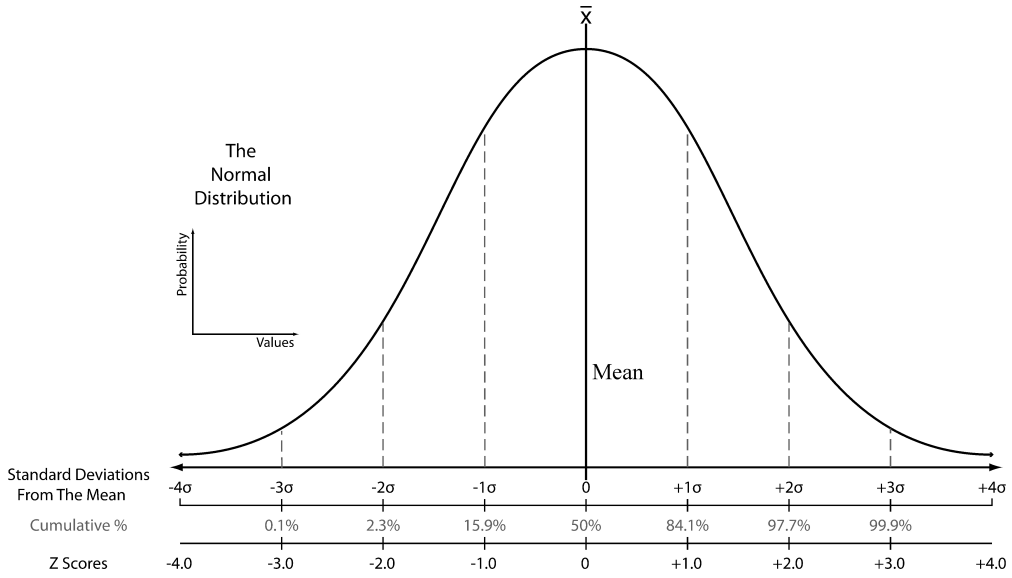



**Figure 10.12:** Comparison of zscore with standard derivation and cumulative percentages.

standard scores, while values below the mean have negative standard scores.

Standard score, which is commonly referred as zscore, is a statistic method to observe the value's distance from the mean. Standard deviation is used as the unit to measure this distance. In general, it measures how many standard deviations away the value is from the mean. For instance, a value will be given a standard score of 1 if it is one standard deviation larger than the mean. Negative standard score will be assigned to those values lower than the mean. Figure 10.12 is the comparison of zscore value with standard derivation and cumulative percentages in a normal distribution.

The equation for zscore calculation is given as following [119]:

$$z = \frac{x - \mu}{\sigma} \tag{10.7.1}$$

in which, $\mu$ is the mean of the population, $\sigma$ is the standard derivation of the

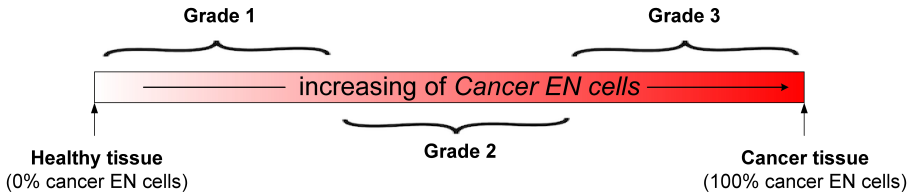
population.

### 10.7.2 Cancer development model (Why zscore analysis)

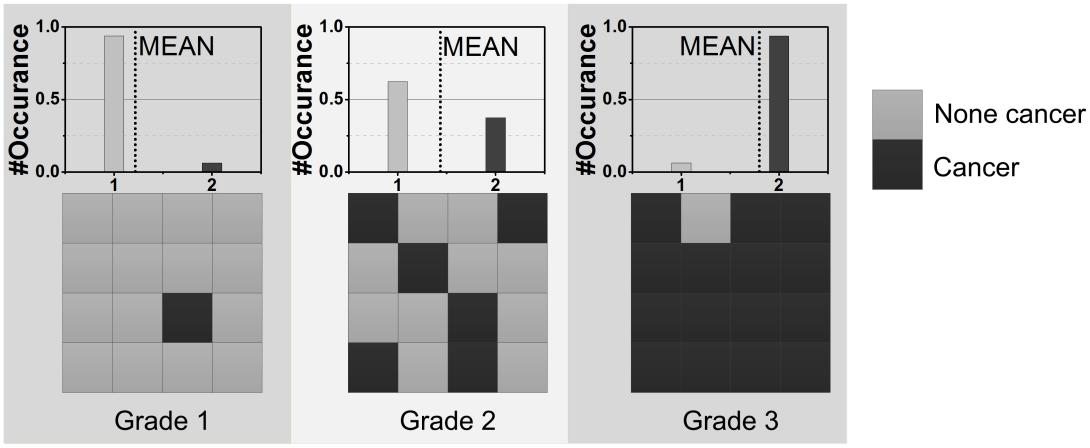
Human breast tumours are histologically complex tissues, containing a variety of cell types in addition to the carcinoma cells [91]. After the background cut, normally, cancer tissues could still have: Lymphocyte cell, Epithelial cell (normal/mitosis/cancer), stroma or blood cells, etc.

A highly simplified tissue model was developed to gain a better understanding of cancer developments and how zscore can be used to reveal the development difference between cancer grades. Previous studies [48, 50, 98] have reported Epithelial (EN) cell to be highly related with cancer developments. To maintain simplicity, only EN cells are considered to be the composition of breast cancer tissue, and only two types: healthy and cancer EN cells are included in the model.

For each tissue, taking the level of cancer development as a spectrum, healthy tissue, in which all the EN cells in the tissue are normal, lies on one end of the spectrum; high-grade cancer, in which almost all the EN cells in the tissue sample are cancer cells, lies on the other end. Low cancer grade (grade one) samples, in which tissue contains only a few cancer cells, are close to the no-cancer end. High cancer (Grade three) samples lie near the cancer end. Having a more robust mixing of cancer and healthy cell, grade two samples are in-between of grade one and three. A detailed illustration of the cancer development spectrum is given in Figure 10.13a.




**(a)** Cancer development spectrum. With healthy tissue on one end, and cancer tissue on the other end. The ratio of cancer EN cells increases as the histological grade of cancer goes up.




**(b)** Schematic diagram of three grades tissue and the histogram analysis. Light grey and black block were used to represent normal and cancer EN cells in the tissue schematic diagram. Assigning 1 to each non-cancer EN block, 2 to reach cancer EN block, histogram analysis of the tissue component is provided on the top of the tissue. The mean of each tissue is marked by dash-line.

**Figure 10.13:** Simplified cancer development model. The histogram analysis in (b) shows that both grade one and three tissue samples have most sub-blocks close to mean, while grade two tissues have a more balanced distribution around the mean. It explains why zscore could be used to identified grade two tissue samples.

Schematic diagrams of tissues from each grade are given in Figure 10.13b, in which, three simplified tissues each from one histological cancer grade are shown. Each tissue is divided into 16 sub-blocks, represented by smaller squares. Sub-blocks with black colour representing cancer EN cells in the tissue, meanwhile, sub-blocks in light grey colour representing normal EN cells. For tissue from grade one, it has 15 light grey and 1 black sub-blocks, indicating 6.25% of its cells are EN cancer cell. Tissues from grade two and three are designed under the same concepts, but with a higher ratio of cancer EN cells.

Given cancer cell value 2 and healthy cell value 1, a numerical model of cancer development can be abstracted from the diagram model. The histogram analyses of each grade sample were conducted, and the mean value of the sample was obtained. The mean and value histogram of different grades are given in Figure 10.13b. For both grade one and three samples, the majority of the sub-blocks are having a value close to the mean. The value distribution among the sample is rather focused on one side (1 for grade one, 2 for grade three), as cancer/normal EN cell distribution across the sample is unbalanced. Grade two, however, has a robust mixing of 1 and 2 values.

Zscore is calculating how many standard derivations away the value is from the mean value. In the case of the cancer development model, taking each sub-block as one individual, most grade one and three samples' individual will have a low standard score as their value close to mean. Grade two sample individuals, however, will have relatively higher zscore values. In other words, the zscore pattern for grade two should be different from zscore patterns for grade one and three.

In real cancer tissue, in terms of the biological components changing, the EN cell cancer development is more than just a yes or no problem. Cells can be in different cancer level. The cancer development model would be far more complex than the one we have in Figure 10.13.

Applying FTIR spectroscopy to a cell is like the numerical process of assigning 1 or 2 to sub-blocks in the cancer development model. Unlike the model, continuous IR absorbance value was assigned to each pixel to describe the various cancer levels. The IR absorption of the pixel is believed containing the cancer information of the cell, which should still be following the zscore changing pattern we described in the cancer development model. From grade one to grade three, cells go from similar (high health cell) to robust mixing of cancer and healthy cells, and then come back to similar (high cancer cell). Grade two, which is in the middle of the cancer development spectrum, has more robust mixing of different level of cancer cells. Grade one and three, meanwhile, due to the highly uniform cells cancel level, less difference between cells can be found. Comparing the zscore value between pixels in the sample could be an effective way to divide grade two samples out.

### 10.7.3 Spectrum zscore analysis

The variation in-between cancer grades can be revealed, by comparing the spectra zscore between tissue sample.

In the cancer development model, one value is directly assigned to each sub-block to represent the cancer level. In the real spectra data cube, taking each

pixel as one sub-block, there is one IR spectrum behind representing the cancer level of that pixel. To make the results comparable to each other, all the dimension reduction method in this chapter is focusing on Amide I peak (wavenumber range 1604 to 1697 $cm^{-1}$). With the spectra resolution, there are still 25 absorbances in each spectrum.

Zscore of each spectrum is calculated using the following steps:

1. Within each sample, taking the absorbance on the same wavenumber as one population sample, the IR absorbance can be divided into 25 population sample sets.

2. Calculate the zscore of each absorbance based on the population data set it is from, using the Equation 10.7.1 on page 255. After this, the spectra in each pixel is transferred into 25 zscore value representing the grading of those 25 absorbances.

3. Average the 25 zscore in each spectrum.

Within the same tissue sample, FTIR spectroscopy data has low sample preparation and data collection variation. Calculating the zscore value within the same sample, inter-sample variation can be excluded.

By comparing the zscore distribution between samples, we are forcing the computational power on biological components difference caused by cancer developments other than the natural uneven distribution. Density distribution method is applied to the zscore value in the next section, to abstract inter-sample difference caused by cancer developments.

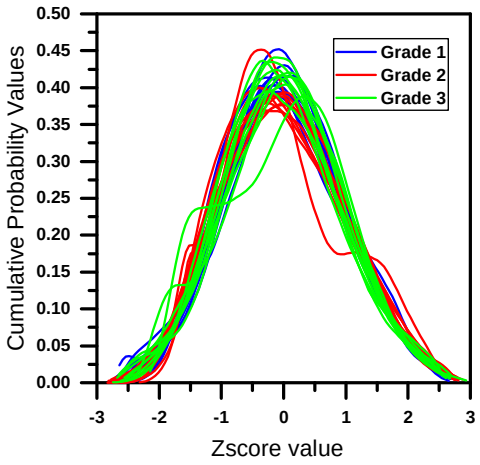
### 10.7.4 Zscore value density distribution

Applying zscore to a spectrum, the whole spectrum can be represented by the mean zscore value, which is taken as the feature value of that spectrum. After the feature extraction process, the 3-D data cube is transferred into a 2-D feature matrix, with each spectrum represented by its own feature value. Bin the feature value within each sample.

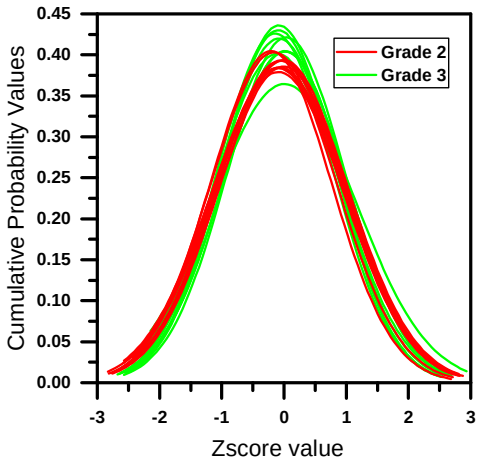
The bin results, the density distribution (Gaussian fitted bin results) and the density distribution function scattering are given in Figure 10.14.

The shape of the density distribution is related to the variation level. The sharper the shape the lower the variation is. The density distribution of the zscore value tends to be a shaper in grade one and three comparing with grades two samples, which means a low uniform level of zscore value, which further indicated a low uniform of cells in middle-grade cancer. This result agreed with the cancer development model we produced in Section 10.7.2. Compared with grade one and three tissue samples, grade two samples, as they are in the middle of cancer development, a more robust mixing of cancer/non-cancer cells is causing the zscore distribution relatively less sharp.

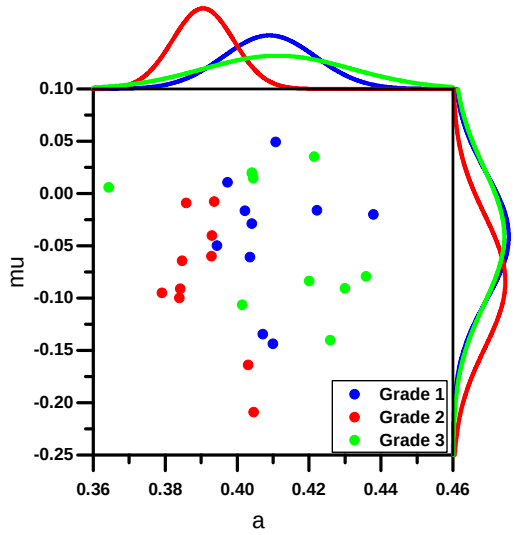
Feeding the density distribution into Gaussian fit to further reduce the dimension of each sample, the density distribution function of each sample is obtained. Parameter $a$ and $\mu$ of each Gaussian fit are chosen as the two variables representing each sample, the scatter plot of the function parameters is given in Figure 10.14c. Figure 10.14d is the same scatter plot with only grade two and three samples. Samples from grade two are divided with grade one and three



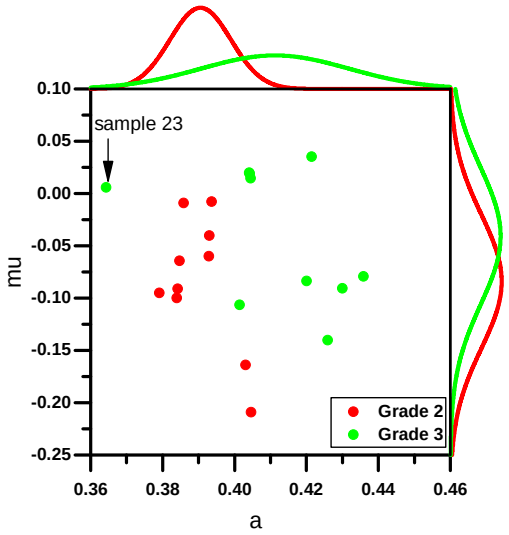

**(a)** Bin plot.




**(b)** Gaussian fitted bin plot.




**(c)** Density distribution function parameter scatter.




**(d)** Density distribution function parameter scatter with grade 2 and 3 samples.

**Figure 10.14:** Zscore analysis results with colour code: blue (grade one), red (grade two) and green (grade three). (c) and (d) is the scatter plot $a$ and $\mu$ with marginal distribution curves. The marginal distribution curves take samples from the same grade as a sub-population. The distribution curve reveals the difference of distribution pattern on each of the dimensions. Samples from grade two are divided with grade one and three tissue samples.

tissue samples.

Outliers were identified in the scatter plot. There are many possible reasons for the outliers: the low data quality (S/N ratio) or the miss-classification by NPI.

Zscore analysis of Amide I peak provides a good example of knowledge-based approach of discovering new knowledge. The hypothesis, grade two tissue sample should have more robust tissue components, are made based on the prior-knowledge, which was evaluated and approved by a specificity designed algorithm, zscore IR spectroscopy data analysis.

## 10.8 Other wavenumber range

Increasing body of evidence indicates that lipid accumulation, especially cholesterol and cholestery1 esters are highly associated with aggressive cancer cells [120, 121].

This chapter, so far, has been focusing on Amide I peak only. There are plenty of other wavenumber range or peaks that can be further explored.

The table provided below has been used and updated by the author throughout his PhD studies. It summarised, if not all, the majority of the IR absorbance fingerprint wavenumber to its biological link, that has been reported in the literature covered by the author.

| **Wavenumber** | **Biological link** |
|---|---|
| 700 | CH lipids [122] |

| | |
|---|---|
| 970 | $PO_3^{2-}$ symmetric, phosphate monoester of phosphorylated proteins and cellular nucleic acids [122] |
| 900-1185 | carbohydrates |
| 1080 | C-O stretch |
| 1084 | $PO_2$- symmetric |
| 1085 & 1241 | antisymmetric and symmetric phosphodiester vibration of nucleic acids |
| 1155 | C-OH groups of serine, threonine, tyrosine in cell proteins [122] |
| 1171 | ester C-O antisymmetric stretch |
| 1230 | O-P-O antisymmetric stretching motions [120] |
| 1236 | antisymmetric vibrations |
| 1238 | $PO_2$- antisymmetric phosphodiester group in nucleic acides and phospholipides [122] |
| 1252 | protein |
| 1300-1400 | various amino acid side chains and fatty acids [123] |
| 1395 | COO- symmetric stretching of amino acids and the symmetric bending mode of the methyl group ($CH_3$) in proteins. |
| 1350-1490 | paraffin [124] |
| 1456 | Lipids, due to antisymmetric vibrations of $CH_3$ |
| 1516 | Tyrosine [114] |
| 1540 & 1650 | Amide I (C=O) and amide II (C-N stretch and H-N-C bend) bands are contributed mainly by protein vibrations [112–116] |
| 1630 | $\beta$-sheet protein secondary structure [114] |

| | |
|---|---|
| 1650 | C=O, Amide I band of tissue and cell proteins [122] |
| 1655 | $\alpha$-helix protein secondary structure [114] |
| 1715 | DNA [125–128] |
| 1740 | >C=O stretching vibrations of the ester carbonyl functional groups in lipids [129] |
| 1741 | phospholipids |
| 1743 | C=O, carbonyl groups of lipids [122] |
| 2800-3010 | lipids, due to their $CH_2$ and $CH_3$ symmetric and antisymmetric vibrations [35, 129, 130] |
| 2872 | symmetric $CH_3$ stretching |
| 2920 | antisymmetric $CH_2$ stretching |
| 2952 | antisymmetric stretching of methyl groups from lipid chains [120] |
| 2957 | antisymmetric $CH_3$ stretching |
| 3280 | H-O-H stretching |

**Table 10.2:** Wavenumber & biochemistry

CHAPTER 11

# Validation using an independent data set

In the previous chapter, we identified the symmetric analysis, and standard score analysis on Amide I peak may be useful for breast cancer prognosis. Further validation, in which the symmetric analysis, zscore analysis, as well as the pre-processing procedure are evaluated on an independent breast cancer FTIR spectroscopy data set, is provided in this Chapter to demonstrate prognostic performance of the methods developed.

## 11.1 Data description

Tissue microarray [131] (TMA) consist of multiple tissue samples of uniform dimensions placed on a single substrate. This arrangement facilitates consistent and convenient processing for all samples after the array is constructed. It is employed not only as high throughput molecular screening tools [132], using, for example, fluorescence techniques [133], but also to address some of the

common challenges in the vibrational spectroscopic analyses of tissue [55].

Increasing demand for high throughput chemical imaging has led to a renewed interest in discrete frequency infrared spectroscopy. Targeting key frequencies instead of acquiring continuous spectra has the potential to dramatically increase throughput [43, 134, 135]. Arguably the most promising contender to date is discrete frequency imaging utilising a tunable, high brightness external cavity infrared quantum cascade laser (QCL) [124]. Exploiting the high brightness of a QCL source [136] enables the optical system to be coupled to an uncooled large area microbolometer, thereby allowing large areas of tissue to be imaged with a single measurement.

Pilling, Henderson and Gardner [124] reported a large study on a breast cancer tissue microarrays (TMA) comprised of 207 different patients. Utilising Spero QCL imaging with continuous spectra acquired between 912 and 1800 $cm^{-1}$, 207 breast cancer cores (15 nonmalignant, 20 grade one, 148 grade two and 7 grade three) FTIR spectroscopy data were collected. The data was made open access online [137].

The TMA contains 16 $\times$ 13 breast tissue cores (with 1 mm diameter, 5 $\mu$m thickness) from different patients. The cores contained 190 cases of invasive ductal carcinoma, 1 mixed lobular and duct carcinoma, 1 mucinous carcinoma, 13 adjacent normal tissue, 3 normal tissue, single core per case. Figure 11.1 is the schematic image explaining the formation of TMA, with each coloured circle representing one tissue core. On the right hand of the image is the H&E stained image of core D1 (research identifier: BR20832 D1), with histology/patient information provided. Information on the pathology classification of each tissue

microarray core can be found in the reference [137].


**Figure 11.1:** Schematic image of the BR20832 TMA. Each coloured circle represent one tissue core. Different colour were used to represent different types of the core, including: Malignant tumour, Malignant tumour (stage 0), Malignant tumour (stage I), Malignant tumour (stage IIIa), Malignant tumour (stage IIIb), Malignant tumour (stage IIa), Malignant tumour (stage IIb), Normal tissue and NAT. On the right hand of the image is the H&E stained image of core D1 (research identifier: BR20832 D1). Detailed information about the patient and the histological result of the core is given. Reproduced from [137].

Infrared chemical images were collected using a QCL infrared microscope (Daylight Solutions Inc., San Diego, CA, United States). FTIR images were acquired in transmission mode using the 4 $\times$ 0.15 NA low magnification objective with a resultant field of view of approximately 2.02 $\times$ 2.02 mm and a corresponding nominal pixel size of approximately 4.2 $\mu$m. Spectra were collected in the spectral range 912-1800 cm$^{-1}$, utilising a step size of 4 cm$^{-1}$ to produce continuous frequency spectra. Each infrared tile consisted of 230400 spectra, was

comprised of 223 data points. Infrared spectra for each biopsy core was extracted from the mosaic as a 313 $\times$ 313 $\times$ 223 data-cube, consisting of 97969 spectra, each with 223 data points. Details about the sample preparation and data collection can be found in reference [124].

This FTIR imaging data set was used to evaluate the performance of the breast cancer FTIR histology method developed. Core row A, B, E and L were selected to include all the grade one and three samples, meanwhile maintain an objective selected, balanced number of grade two samples.

## 11.2 Pre-processing

Pilling, Henderson and Gardner [124] tested the quality of the spectra to remove data obtained from areas with little or no tissue using the height of the amide I band with spectra having absorbance between 0.1 and 2 being retained. Principal component-based noise reduction was used to improve signal-to-noise with the first 40 PCs being retained. Spectra were truncated between 1000 and 1800 $cm^{-1}$, and the region describing the absorption bands of wax (1350-1490 $cm^{-1}$) were removed. Each spectrum was then vector normalised to correct for different thicknesses of tissue and finally converted to its first derivative while performing Savitzky-Golay smoothing using a window size of nine data points.

Savitzky-Golay smoothing using a window size of nine data points was applied on spectra data before further analysis. Spectra truncation is not necessary for the QLC spectrum, as spectral range 912-1800 $cm^{-1}$ is fit for data handling, and the dimension reduction methods developed can identify Amide I peak and

input that into algorithm automatically. The background cut reported in 9.2.2 is based on the standard derivation difference of absorbance value between 1800-3000 $cm^{-1}$. This wavenumber range is not included when the QLC spectra were collected. The background cut for QLC data set was conducted based on the absorbance value of Amide I peak (height of the amide I band) which is the same method reported in the paper. A background ground cut-off points of 0.1 is used to identify those empty pixels. Same as reported in the paper, spectra with Amide I peak height exceeds 2 were cut off from the dataset before further analysis was applied.

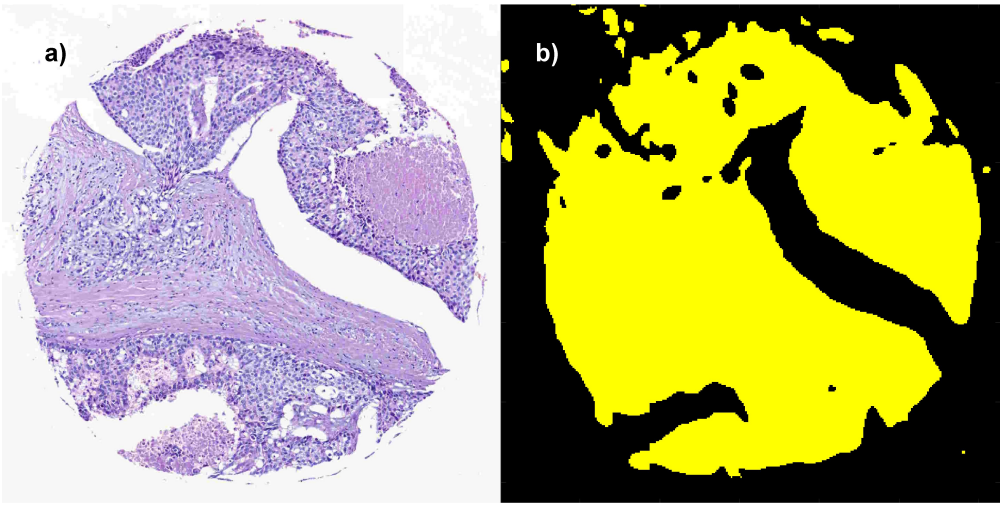


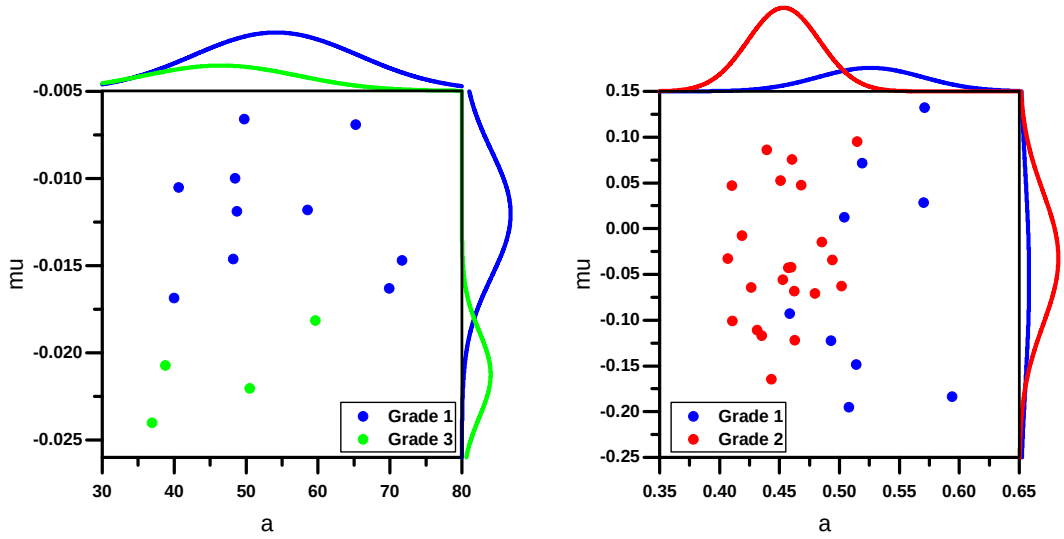

**Figure 11.2:** (a) Brightfield image of H&E stained serial section for core A12 and (b) pseudo-colour image of the background identification using yellow (tissue structure) and black (empty background).

Figure 11.2a shows a high-resolution bright-field image of a mixed core (core A12) consisting primarily of epithelium and stroma. Comparison of the background identified pseudo-colour image (Figure 11.2b, yellow $\longrightarrow$ tissue structure and black $\longrightarrow$ empty background) to the bright-field image of the H&E stained section illustrates that there is a good agreement on the background

identification.

## 11.3 Results

Symmetric analysis and zscore analysis were applied on the QLC data set. The scatter images of the fitted density distribution function parameters are given in Figure 11.3. An outlier cut was applied to cut off those core samples with low representativeness.




**(a)** Symmetric analysis validation on QLC data set.

**(b)** Zscore analysis validation on QLC data set.

**Figure 11.3:** Validation of breast cancer prognosis methods on an independent QLC FTIR spectroscopy data set with colour code: blue (grade one), red (grade two) and green (grade three). The distribution curve reveals the difference of distribution pattern on each of the dimensions.

The symmetric analysis result shows a clear cluster between grade one and three samples. Compared with grade one samples (blue dots in Figure 11.3a), grade three samples (green dots in Figure 11.3a) have low $\mu$ value in the fitted

density distribution function, which means, grade three samples tend to have lower average symmetric value (definition is given in section 10.6.4) over all the spectra selected. The result was found to be in good agreement with the 30 tissue sample symmetric analysis results in Figure 10.11 on page 252, which grade three sample tends to have a lower $\mu$ parameter for the fitted density distribution function. The scatter difference on the $a$ parameter observed in the 30 tissue sample is not observed in the QCM dataset. That might due to the imaged tissue region is much smaller in the QCM tissue cores.

The zscore analysis result shows a clear cluster between grade one and three samples. Compared with grade two samples (red dots in Figure 11.3b), grade one samples (blue dots in Figure 11.3b) have higher $a$ value in the fitted density distribution function, which means, grade one samples tend to have less robust zscore value distribution over all the spectra selected. The result was found to be in good agreement with the 30 tissue sample zscore analysis result in Figure 10.14 on page 262, which grade one sample tends to have a higher $a$ parameter for the fitted density distribution function.

Both the symmetric analysis and zscore analysis results show a clear cluster of samples from different histopathology grade. The scatter patterns of both the symmetric analyses and the zscore analyses fit well with the results observed in Chapter 10. The multi-stage data analysis algorithm developed can provide statistical control over the breast cancer classification process and produce a precise cancer prognosis on tissue sample level. With a simple structure and clear data handing procedure, those breast cancer algorithms have good generalisations.

Performance difference can still be observed. In the symmetric analysis result of the 30 sample dataset, grade three samples tend to have higher $a$ parameter compared with grade one samples, which means grade three samples tend to have less variation within the same sample, because the density distributions of symmetric value have shaper Gaussian shape (higher $a$). This pattern is not observed in the QLC dataset results. In Figure 11.3a a clear cluster can be observed on the $\mu$ dimension. The $a$ dimension provide limited dividing information between grade one and three samples. The distribution curve, blue and green line on the $a$ dimension reveals that grade three samples, in general, have lower $a$ value. This might due to the QLC dataset collected significant smaller tissue region for FTIR imaging. Each TMA core has the area of 0.79 $mm^2$, while each tissue sample in the 30 sample dataset has the area of 7.84 $mm^2$. The sample size is much larger in the 30 sample dataset.

The confidence of this evaluation is limited by the unbalanced number between grade one, two and three samples. The TMA contains 20 grade one, 148 grade two and 7 grade three samples. The representativeness of those 7 grade three samples is not high, compared with the 148 available grade two samples. Further evaluation of the method on a bigger and more balanced dataset is necessary. We are currently working on the access of such dataset.

CHAPTER 12

# Conclusion

This study seeks to develop object and transparent spectra data analysis technique to address this identified research gap. A multi-stage data analysis algorithm developed in this work, can provide statistical control over the breast cancer classification process and produce an accurate cancer prognosis.

In Chapter 9, standard pre-processing methods were discussed and selected in an objective way. Pre-processing methods that would make the transferred data lost its physical meanings were discarded. Background and detector edge pixels were identified and cut off from the data cube. Minimum manipulation of the original data was maintained throughout this chapter to provide objectiveness and statistic confidence for future analysis.

In Chapter 10, density distribution function, a statistical approach of abstracting cancer relevant inter-sample difference, was proposed. PCA, as the most-used dimension reduction method was tried. Compared with PCA results, symmetric analysis and zscore analysis, as a dimension reduction technique, can provide better performance in terms of separating tissue samples from differ-

ent grades.

Throughout the breast cancer work in this thesis, NPI grading results were used as the true value that we were trying to match with, using the FTIR spectroscopy results. As introduced in Chapter 8, the NPI grading result is based on a numerical prognostic index. It categorised the numerical prognosis index into three different grade (grade one two three), using the cut off value 3.4 and 5.4. It might be better to match the symmetric analysis or zscore analysis result directly with the numerical prognostic index, as a numerical system describing cancer stage of a patient is preferable than three stages system.

## 12.1 Where we are heading to

### 12.1.1 Further improvements of the prognosis system

Nuclei are one of the most important histologic primitives in cancer prognosis. Cell nuclei feature such as size, texture, shape, and other chemical components change along with the development of cancer. Classification and grading of cancer are highly dependent on the nuclei area. Among the different types of nuclei, two types are usually the object of particular interest: lymphocyte and epithelial nuclei. Lymphocyte nuclei (LN) has regular shape and normally is smaller than epithelial nuclei (EN) (see Figure 9.1a on page 203). Normal EN has nearly uniform chromatin distribution with smooth boundary (see Figure 9.1b). In high-grade cancer tissue, EN is larger in size. They may have heterogeneous chromatin distribution, irregular boundaries and clearly visible nucleoli compared to normal EN (see Figure 9.1e) [34].

Identifying high-grade cancer EN is essential for cancer grading. To achieve this, in the traditional histopathology approach, hematoxylin is used to visualise the nuclei structure. Histopathologists pick out the high-grade cancer EN with the help of microscopy. The chemical components change in the nuclei as the development of cancer make it possible to detect high-grade cancer EN based on FTIR spectroscopy data. Histological identification and segmentation steps can be added to abstract pixels with high cancer representatives to improve the performance of the prognosis system developed in Chapter 9 and 10.

### 12.1.2 Personalised medicine

Personalised medicine is one of the strategic research goals of breast cancer research. The current NPI system is dividing breast cancer patients into three groups: grade one, two and three to describe the medical condition of the patients. Considering the complexity of cancer and the patients' individual difference, it is far away from enough. The robust information of IR spectrum provides the possibility of tailor treatments of breast cancer according to the biological characteristics of the cancer tissue and the specific needs of the individual.

### 12.1.3 Academic pathology

Academic pathology can be achieved using FTIR imaging technique. Breast cancer is not a single disease, but a combination of diseases [36]. Based on the estrogen receptor, breast cancer can be divided into two types: ER-positive and

ER-negative. Majority of breast cancer patients have ER-postive cancer. Based on gene expression studies, breast cancer can be divided into five subtypes: luminal A, luminal B, ERBB2$^+$, normal breast-like and basal-live (triple-negative). Luminal A and B are derived from ER-positive type while ERBB2$^+$, normal breast-like and basal-live are derived from ER-negative type [138, 139].

30% cases of breast cancer do not respond well to current treatment methods. Patients within the treatable group (ER-positive), nearly a quarter of them developed into an aggressive type as the cancer developments. There is an urgent need to develop better understanding and treatment to this aggressive type of breast cancer. The concept that protein is highly related to the aggressive breast cancer type is demonstrated by increasing number of researches [140, 141] The unique fingerprint signature of the secondary structure of a protein in IR spectrum could provide a reliable way of understanding the cancer development.

### 12.1.4 Other application

Further application of the spectra analysis method is possible. Infrared and Raman spectroscopy are complementary techniques, which can provide subtly different information about a sample [142]. The methods for spectra evaluation and analysis are similar [143]. The multivariate spectra analysis algorithm developed in this work can be widely applied to bio-spectroscopy researches.

# Bibliography


[1] Ahmedin Jemal et al. "Global cancer statistics". In: *CA. Cancer J. Clin.* 61.2 (2011), pp. 69–90.

[2] Colin Mathers, Doris Ma Fat and Jan Ties Boerma. *The global burden of disease (2004 update)*. World Health Organization, 2008.

[3] U. K. Breast Cancer. 2015. URL: `http://www.breastcanceruk.org.uk/` (visited on 17/10/2015).

[4] Marcus H. Galea et al. "The Nottingham Prognostic Index in primary breast cancer". In: *Breast Cancer Res. Treat.* 22.3 (1992), pp. 207–219.

[5] Heinz Fabian et al. "A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts". In: *Biospectroscopy* 1.1 (1995), pp. 37–45.

[6] Michael Jackson et al. "Classification of breast tumors by grade and steroid receptor status using pattern recognition analysis of infrared spectra." In: *Cancer Detect. Prev.* 23.3 (1998), pp. 245–253.

[7] Frances Nell Pounder, Rohith Krishna Reddy and Rohit Bhargava. "Development of a practical spatial-spectral analysis protocol for breast histopathology

using Fourier transform infrared spectroscopic imaging". In: *Faraday Discuss.* (2016).

[8] Rainer Eckel et al. "Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue". In: *Vib. Spectrosc.* 27.2 (2001), pp. 165–173.

[9] Yun Xiang Ci et al. "Fourier transform infrared spectroscopic characterization of human breast tissue: implications for breast cancer diagnosis". In: *Appl. Spectrosc.* 53.3 (1999), pp. 312–315.

[10] Wenyue Yang et al. "In situ evaluation of breast cancer cell growth with 3D ATR-FTIR spectroscopy". In: *Vib. Spectrosc.* 49.1 (2009), pp. 64–67.

[11] Audrey Benard et al. "Infrared imaging in breast cancer: automated tissue component recognition and spectral characterization of breast cancer cells as well as the tumor microenvironment". In: *Analyst* 139.5 (2014), pp. 1044–1056.

[12] Benjamin Bird et al. "Detection of breast micro-metastases in axillary lymph nodes by infrared micro-spectral imaging". In: *Analyst* 134.6 (2009), pp. 1067–1076.

[13] Angela Kallenbach-Thieltges et al. "Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections". In: *J. Biophotonics* 6.1 (2013), pp. 88–100.

[14] Benjamin Bird et al. "Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancer". In: *Lab. Invest.* 92.9 (2012), pp. 1358–1373.

[15] Matthew J. German et al. "Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell". In: *Biophys. J.* 90.10 (2006), pp. 3783–3795.

[16] Peter R. Griffiths and James A. De Haseth. *Fourier transform infrared spectrometry*. John Wiley & Sons, 2007.

[17] D. Michael Byler and Heino Susi. "Examination of the secondary structure of proteins by deconvolved FTIR spectra". In: *Biopolymers* 25.3 (1986), pp. 469–487.

[18] Michael Jackson and Henry H. Mantsch. "The medical challenge to infrared spectroscopy". In: *J. Mol. Struct.* 408 (1997), pp. 105–111.

[19] Dieter Naumann, Dieter Helm and Harald Labischinski. "Microbiological characterizations by FTIR spectroscopy." In: *Nature* 351.6321 (1991), pp. 81–82.

[20] P. T. Wong et al. "Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis". In: *PNAS* 88.24 (1991), pp. 10988–10992.

[21] Donald C. Malins, Nayak L. Polissar and Sandra J. Gunselman. "Models of DNA structure achieve almost perfect discrimination between normal prostate, benign prostatic hyperplasia (BPH), and adenocarcinoma and have a high potential for predicting BPH and prostate cancer". In: *PNAS* 94.1 (1997), pp. 259–264.

[22] Paul G. L. Andrus and Robert D. Strickland. "Cancer grading by Fourier transform infrared spectroscopy". In: *Biospectroscopy* 4.1 (1998), pp. 37–46.

[23] H. M. Heise, L. Küpper and L. N. Butvina. "Bio-analytical applications of mid-infrared spectroscopy using silver halide fiber-optic probes". In: *Spectrochimica. Acta B* 57.10 (2002), pp. 1649–1663.

[24] Kaustuv Das et al. "FTIR of touch imprint cytology: a novel tissue diagnostic technique". In: *J. Photochem. Photobiol. B, Biol.* 92.3 (2008), pp. 160–164.

[25] Mark J. Tobin et al. “FTIR spectroscopy of single live cells in aqueous media by synchrotron IR microscopy using microfabricated sample holders”. In: *Vib. Spectrosc.* 53.1 (2010), pp. 34–38.

[26] Julian Ollesch et al. “FTIR spectroscopy of biofluids revisited: an automated approach to spectral biomarker identification”. In: *Analyst* 138.14 (2013), pp. 4092–4102.

[27] Benjamin L. Bird. “FTIR imaging: a route toward automated histopathology”. PhD thesis. University of Nottingham, 2006.

[28] R. Barer. “Aspects of ultra-violet and infra-red microspectrography with the Burch reflecting microscope”. In: *Discuss Faraday Soc* 9 (1950), pp. 369–378.

[29] Vincent J. Coates, Abe Offner and E. H. Siegler. “Design and performance of an infrared microscope attachment”. In: *JOSA* 43.11 (1953), pp. 984–989.

[30] Albert A. Michelson. “XXXVIII. On the application of interference-methods to spectroscopic measurements.—I”. In: *Lond.Edinb.Dubl.Phil.Mag.* 31.191 (1891), pp. 338–346.

[31] Rohit Bhargava, Bentley G. Wall and Jack L. Koenig. “Comparison of the FT-IR mapping and imaging techniques applied to polymeric systems”. In: *Appl. Spectrosc.* 54.4 (2000), pp. 470–479.

[32] Bayden R. Wood. “The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues”. In: *Chem. Soc. Rev.* 45.7 (2016), pp. 1980–1998.

[33] Arianna Ricci et al. “Application of Fourier Transform Infrared (FTIR) Spectroscopy in the Characterization of Tannins”. In: *Appl. Spectrosc. Rev.* just-accepted (2015), pp. 00–00.

[34] Humayun Irshad et al. "Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential". In: *IEEE Rev. Biomed. Eng.* 7 (2014), pp. 97–114.

[35] Zanyar Movasaghi, Shazza Rehman and Dr Ihtesham ur Rehman. "Fourier transform infrared (FTIR) spectroscopy of biological tissues". In: *Appl. Spectrosc. Rev.* 43.2 (2008), pp. 134–179.

[36] Tracy Vargo-Gogola and Jeffrey M. Rosen. "Modelling breast cancer: one size does not fit all". In: *Nat. Rev. Cancer* 7.9 (2007), p. 659.

[37] Ira W. Levin and Rohit Bhargava. "FOURIER TRANSFORM INFRARED VIBRATIONAL SPECTROSCOPIC IMAGING: Integrating Microscopy and Molecular Recognition*". In: *Annu. Rev. Phys. Chem.* 56 (2005), pp. 429–474.

[38] G. L. Carr. "Resolution limits for infrared microspectroscopy explored with synchrotron radiation". In: *Rev. Sci. Instrum.* 72.3 (2001), pp. 1613–1619.

[39] Matthew J. Baker et al. "Using Fourier transform IR spectroscopy to analyze biological materials". In: *Nat. Protoc.* 9.8 (2014), pp. 1771–1791.

[40] Tomasz P. Wrobel et al. "Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy of a single endothelial cell". In: *Analyst* 137.18 (2012), pp. 4135–4139.

[41] Kamila Kochan et al. "Comparison of FTIR transmission and transfection substrates for canine liver cancer detection". In: *Analyst* 140.7 (2015), pp. 2402–2411.

[42] Julie Cao et al. "Fourier transform infrared microspectroscopy reveals that tissue culture conditions affect the macromolecular phenotype of human embryonic stem cells". In: *Analyst* 138.14 (2013), pp. 4147–4160.

[43] Paul Bassan et al. “Transmission FT-IR chemical imaging on glass substrates: applications in infrared spectral histopathology”. In: *Anal. Chem.* 86.3 (2014), pp. 1648–1653.

[44] Michael J. Pilling et al. “Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation”. In: *Analyst* (2017).

[45] Rohit Bhargava, Shi-Qing Wang and Jack L. Koenig. “Processing FT-IR imaging data for morphology visualization”. In: *Appl. Spectrosc.* 54.11 (2000), pp. 1690–1706.

[46] Frances N. Keith, Rohith K. Reddy and Rohit Bhargava. “Practical protocols for fast histopathology by Fourier transform infrared spectroscopic imaging”. In: *Int. Biomedical. Optics.* Vol. 6853. SPIE. 2008, pp. 06–1.

[47] F. Nell Pounder et al. “Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging”. In: *P. Soc. Photo-opt. Ins.* 10 (2009), p. 26.

[48] Rohit Bhargava. “Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology”. In: *Anal. Bioanal. Chem.* 389.4 (2007), pp. 1155–1169.

[49] Joëlle De Meutter et al. “Infrared imaging of high density protein arrays”. In: *Analyst* (2017).

[50] Michael J. Nasse et al. “High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams”. In: *Nat. Methods* 8.5 (2011), pp. 413–416.

[51] D. Petersen et al. “Virtual staining of colon cancer tissue by label-free Raman micro-spectroscopy”. In: *Analyst* (2017).

[52] A. Kohler et al. "Estimating and correcting Mie scattering in synchrotron-based microscopic Fourier transform infrared spectra by extended multiplicative signal correction". In: *Appl. Spectrosc.* 62.3 (2008), pp. 259–266.

[53] J. L. Haybittle et al. "A prognostic index in primary breast cancer." In: *Br. J. Cancer* 45.3 (1982), p. 361.

[54] Halina Abramczyk et al. "The label-free Raman imaging of human breast cancer". In: *J. Mol. Liq.* 164.1 (2011), pp. 123–131.

[55] Rohit Bhargava et al. "High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data". In: *BBA-Biomembranes* 1758.7 (2006), pp. 830–845.

[56] Andrew H. Fischer et al. "Hematoxylin and eosin staining of tissue and cell sections". In: *CSH Protocols* 2008.5 (2008), pp. 49–86.

[57] Christopher S. Stapleton. "Spectroscopic studies of pharmaceutical and biological materials". PhD thesis. University of Nottingham, 2013.

[58] C. W. Elston and I. O. Ellis. "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up". In: *Histopathology* 19.5 (1991), pp. 403–410.

[59] Marko Simunovic et al. "A snapshot of waiting times for cancer surgery provided by surgeons affiliated with regional cancer centres in Ontario". In: *Can. Med. Assoc. J.* 165.4 (2001), pp. 421–425.

[60] Christos Sotiriou et al. "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis". In: *J. Natl. Cancer Inst.* 98.4 (2006), pp. 262–272.

[61] Patrick L. Fitzgibbons et al. "Prognostic factors in breast cancer: College of American Pathologists Consensus Statement 1999." In: *Arch. Pathol. Lab. Med.* 124.7 (2000), pp. 966–978.

[62] S. Eva Singletary et al. "Revision of the American Joint Committee on Cancer staging system for breast cancer". In: *J. Clin. Oncol.* 20.17 (2002), pp. 3628–3636.

[63] H. Fabian et al. "Mid-IR microspectroscopic imaging of breast tumor tissue sections". In: *Biopolymers* 67.4-5 (2002), pp. 354–357.

[64] Heinz Fabian et al. "Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy". In: *BBA-Biomembranes* 1758.7 (2006), pp. 874–882.

[65] S. Kumar et al. "Change in the microenvironment of breast cancer studied by FTIR imaging". In: *Analyst* 138.14 (2013), pp. 4058–4065.

[66] Sebastian Berisha et al. "Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks". In: *Analyst* (2018).

[67] Benjamin Bird et al. "Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology". In: *BMC Clin. Pathol.* 8.1 (2008), p. 8.

[68] Xiao Ying Wang. "Fuzzy clustering in the analysis of Fourier transform infrared spectra for cancer diagnosis". PhD thesis. Citeseer, 2006.

[69] Shabbar Naqvi. "Modelling FTIR spectral sata with Type-I and Type-II fuzzy sets for breast cancer grading". PhD thesis. University of Nottingham, 2014.

[70] Júlio Trevisan et al. "Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives". In: *Analyst* 137.14 (2012), pp. 3202–3215.

[71] Max Diem et al. “Molecular pathology via IR and Raman spectral imaging”. In: *J. Biophotonics* 6.11-12 (2013), pp. 855–886.

[72] Benjamin Bird and Jeremy Rowlette. “High Definition Infrared Chemical Imaging of Colorectal Tissue using a Spero QCL Microscope”. In: *Analyst* (2017).

[73] Daniel C. Fernandez et al. “Infrared spectroscopic imaging for histopathologic recognition”. In: *Nat. Biotechnol.* 23.4 (2005), pp. 469–474.

[74] Peter Lasch et al. “Imaging of colorectal adenocarcinoma using FTIR microspectroscopy and cluster analysis”. In: *BBA-Mol. Basis.Dis.* 1688.2 (2004), pp. 176–186.

[75] Caryn Hughes et al. “FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder”. In: *J. Biophotonics* 6.1 (2013), pp. 73–87.

[76] Matthew James Baker et al. “An investigation of the RWPE prostate derived family of cell lines using FTIR spectroscopy”. In: *Analyst* 135.5 (2010), pp. 887–894.

[77] Harald Martens et al. “Pre-processing in biochemometrics: correction for path-length and temperature effects of water in FTIR bio-spectroscopy by EMSC”. In: *J. Chemom.* 20.8-10 (2006), pp. 402–417.

[78] Roger M. Jarvis and Royston Goodacre. “Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data”. In: *Bioinformatics* 21.7 (2005), pp. 860–868.

[79] Richard O. Duda, Peter E. Hart and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[80] David H. Wolpert. “Stacked generalization”. In: *Neural Networks* 5.2 (1992), pp. 241–259.

[81] Yanfeng Gao et al. “Fourier transform infrared microspectroscopy monitoring of 5-fluorouracil-induced apoptosis in SW620 colon cancer cells”. In: *Mol. Med. Rep.* 11.4 (2015), pp. 2585–2591.

[82] Jemma G. Kelly et al. “Robust classification of low-grade cervical cytology following analysis with ATR-FTIR spectroscopy and subsequent application of self-learning classifier eClass”. In: *Anal. Bioanal. Chem.* 398.5 (2010), pp. 2191–2201.

[83] Seng Khoon Teh et al. “Spectroscopic diagnosis of laryngeal carcinoma using near-infrared Raman spectroscopy and random recursive partitioning ensemble techniques”. In: *Analyst* 134.6 (2009), pp. 1232–1239.

[84] Peter Lasch et al. “Imaging of human colon carcinoma thin sections by FTIR microspectrometry”. In: *Int. Biomedical. Optics.* SPIE. 1998, pp. 187–198.

[85] C. Beleites et al. “Classification of human gliomas by infrared imaging spectroscopy and chemometric image processing”. In: *Vib. Spectrosc.* 38.1 (2005), pp. 143–149.

[86] Rajmund L. Somorjai et al. “A data-driven, flexible machine learning strategy for the classification of biomedical data”. In: *Artificial intelligence methods and tools for systems biology*. Springer, 2004, pp. 67–85.

[87] Dieter Naumann. “FTIR spectroscopy of microorganisms at the Robert Koch Institute: experiences gained during a successful project”. In: *Int. Biomedical. Optics.* Vol. 530. SPIE. 2008, pp. 68–68.

[88] Francis L. Martin et al. “Distinguishing cell types or populations based on the computational analysis of their infrared spectra”. In: *Nat. Protoc.* 5.11 (2010), pp. 1748–1760.

[89] Douglas B. Kell. "Genotype–phenotype mapping: genes as computer programs". In: *Trends Genet.* 18.11 (2002), pp. 555–559.

[90] Royston Goodacre et al. "Metabolomics by numbers: acquiring and understanding global metabolite data". In: *Trends Biotechnol.* 22.5 (2004), pp. 245–252.

[91] Lone Ronnov-Jessen, Ole W. Petersen and Mina J. Bissell. "Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction". In: *Physiol. Rev.* 76.1 (1996), pp. 69–125.

[92] Paul Bassan et al. "Resonant Mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact'". In: *Analyst* 134.8 (2009), pp. 1586–1593.

[93] Paul Bassan et al. "Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples". In: *Analyst* 135.2 (2010), pp. 268–277.

[94] C. Kuepper et al. "Label-free classification of colon cancer grading using infrared spectral histopathology". In: *Faraday Discuss.* 187 (2016), pp. 105–118.

[95] O. J. Old et al. "Rapid infrared mapping for highly accurate automated histology in Barrett's oesophagus". In: *Analyst* (2017).

[96] Melissa Romeo and Max Diem. "Correction of dispersive line shape artifact observed in diffuse reflection infrared spectroscopy and absorption/reflection (transflection) infrared micro-spectroscopy". In: *Vib. Spectrosc.* 38.1 (2005), pp. 129–132.

[97] Claus Kuepper et al. "Quantum Cascade Laser-Based Infrared Microscopy for Label-Free and Automated Cancer Classification in Tissue Sections". In: *Sci. Rep.* 8.1 (2018), p. 7717.

[98] Rohit Bhargava and Anant Madabhushi. "Emerging themes in image informatics and molecular analysis for digital pathology". In: *Annu. Rev. Biomed. Eng.* 18 (2016), pp. 387–412.

[99] John Jacob Fox and A. E. Martin. "Investigations of infra-red spectra. Determination of CH frequencies (˜ 3000 cm.-1) in paraffins and olefins, with some observations on "polythenes"". In: *Proc. R. Soc. Lond. A* 175.961 (1940), pp. 208–233.

[100] L. Kellner. "The Vibrations of an Infinitely Long Chain of CH2-Groups and the Infra-Red Spectrum of Polythene". In: *Proc. Phys. Soc. London, Sect. A* 64.6 (1951), p. 521.

[101] S. C. Y. L. Krimm, C. Y. Liang and G. B. B. M. Sutherland. "Infrared spectra of high polymers. II. Polyethylene". In: *J. Chem. Phys.* 25.3 (1956), pp. 549–562.

[102] Prabuddha Mukherjee et al. "Relating Post-yield Mechanical Behavior in Polyethylenes to Spatially Varying Molecular Deformation Using Infrared Spectroscopic Imaging: Homopolymers". In: *Macromolecules* (2018).

[103] Witold K. Surewicz, Henry H. Mantsch and Dennis Chapman. "Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment". In: *Biochemistry* 32.2 (1993), pp. 389–394.

[104] Franck Bonnier et al. "Ultra-filtration of human serum for improved quantitative analysis of low molecular weight biomarkers using ATR-IR spectroscopy". In: *Analyst* 142.8 (2017), pp. 1285–1298.

[105] J. Michael Hollas. *Modern spectroscopy*. John Wiley & Sons, 2004.

[106] Lucia Mutihac and Radu Mutihac. "Mining in chemometrics". In: *Anal. Chim. Acta.* 612.1 (2008), pp. 1–18.

[107] Jeffrey T. Leek et al. "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nat. Rev. Genet.* 11.10 (2010), pp. 733–739.

[108] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

[109] Hajime Torii and Mitsuo Tasumi. "Model calculations on the amide-I infrared bands of globular proteins". In: *J. CHEM. PHYS.* 96.5 (1992), pp. 3379–3387.

[110] Cyril Petibois and Gérard Déléris. "Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology". In: *Trends Biotechnol.* 24.10 (2006), pp. 455–462.

[111] Véronique Cabiaux et al. "Secondary structure of diphtheria toxin and its fragments interacting with acidic liposomes studied by polarized infrared spectroscopy." In: *J. Biol. Chem.* 264.9 (1989), pp. 4928–4938.

[112] Eoghan Ó. Faoláin et al. "Raman spectroscopic evaluation of efficacy of current paraffin wax section dewaxing agents". In: *J. Histochem. Cytochem.* 53.1 (2005), pp. 121–129.

[113] D. Naumaim. "Infrared and NIR Raman spectroscopy in medical microbiology". In: *Proceedings of SPIE*. Vol. 3257. 1998, pp. 245–257.

[114] Erik Goormaghtigh, Jean-Marie Ruysschaert and Vincent Raussens. "Evaluation of the information content in infrared spectra for protein secondary structure determination". In: *Biophys J* 90.8 (2006), pp. 2946–2957.

[115] Erik Goormaghtigh et al. "Protein secondary structure content in solution, films and tissues: redundancy and complementarity of the information content in circular dichroism, transmission and ATR FTIR spectra". In: *BBA-Proteins Ptoteom* 1794.9 (2009), pp. 1332–1343.

[116] Harmen H. J. De Jongh, Erik Goormaghtigh and Jean-Marie Ruysschaert. "The different molar absorptivities of the secondary structure types in the amide I region: an attenuated total reflection infrared study on globular proteins". In: *Anal. Biochem.* 242.1 (1996), pp. 95–103.

[117] M. M. Coleman and J. Zarian. "Fourier-transform infrared studies of polymer blends. II. Poly (caprolactone)–poly (vinyl chloride) system". In: *J. Polym. Sci., Part B: Polym. Phys.* 17.5 (1979), pp. 837–850.

[118] S. Yu Venyaminov and N. N. Kalnin. "Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. I. Spectral parameters of amino acid residue absorption bands". In: *Biopolymers* 30.13-14 (1990), pp. 1243–1257.

[119] Dennis Zill, Warren S. Wright and Michael R. Cullen. *Advanced engineering mathematics*. Jones & Bartlett Learning, 2011.

[120] Ewelina Lipiec et al. "Nanoscale investigation into the cellular response of glioblastoma cells exposed to protons". In: *Anal. Chem.* (2018).

[121] S. Beloribi-Djefaflia, S. Vasseur and F. Guillaumond. "Lipid metabolic reprogramming in cancer cells". In: *Oncogenesis* 5.1 (2016), e189.

[122] Marc Meurens et al. "Breast cancer detection by Fourier transform infrared spectrometry". In: *Vib. Spectrosc.* 10.2 (1996), pp. 341–346.

[123] Allison Derenne, Régis Gasper and Erik Goormaghtigh. "The FTIR spectrum of prostate cancer cells allows the classification of anticancer drugs according to their mode of action". In: *Analyst* 136.6 (2011), pp. 1134–1141.

[124] Michael J. Pilling, Alex Henderson and Peter Gardner. "Quantum Cascade Laser Spectral Histopathology: Breast Cancer Diagnostics Using High Throughput Chemical Imaging". In: *Anal. Chem.* 89.14 (2017), pp. 7348–7355.

[125] Donna R. Whelan et al. "Quantification of DNA in simple eukaryotic cells using Fourier transform infrared spectroscopy". In: *J Biophotonics* 6.10 (2013), pp. 775–784.

[126] Donna R. Whelan et al. "Monitoring the reversible B to A-like transition of DNA in eukaryotic cells using Fourier transform infrared spectroscopy". In: *Nucleic Acids Res.* 39.13 (2011), pp. 5439–5448.

[127] Donna R. Whelan et al. "Detection of an en masse and reversible B-to A-DNA conformational transition in prokaryotes in response to desiccation". In: *J. Royal Soc. Interface* 11.97 (2014), p. 20140454.

[128] Donna R. Whelan et al. "Synchrotron Fourier transform infrared (FTIR) analysis of single living cells progressing through the cell cycle". In: *Analyst* 138.14 (2013), pp. 3891–3899.

[129] Urs Peter Fringeli and Hs H. Günthard. "Infrared membrane spectroscopy". In: *Membrane spectroscopy*. Springer, 1981, pp. 270–332.

[130] Allison Derenne, Olivier Vandersleyen and Erik Goormaghtigh. "Lipid quantification method using FTIR spectroscopy applied on cancer cell extracts". In: *BBA-Mol. Cell Res.* 1841.8 (2014), pp. 1200–1209.

[131] Juha Kononen et al. "Tissue microarrays for high-throughput molecular profiling of tumor specimens". In: *Nat. Med.* 4.7 (1998), p. 844.

[132] Till Braunschweig, Joon-Yong Chung and S. M. Hewitt. "Tissue microarrays: bridging the gap between research and the clinic". In: *Expert Rev. Proteomics* 2.3 (2005), pp. 325–336.

[133] Robert L. Camp, Gina G. Chung and D. L. Rimm. "Automated subcellular localization and quantification of protein expression in tissue microarrays". In: *Nat. Med.* 8.11 (2002), p. 1323.

[134] Gavin Rhys Lloyd and Nicholas Stone. "Method for identification of spectral targets in discrete frequency infrared spectroscopy for clinical diagnostics". In: *Appl. Spectrosc.* 69.9 (2015), pp. 1066–1073.

[135] Shachi Mittal et al. "A four class model for digital breast histopathology using high-definition Fourier transform infrared (FT-IR) spectroscopic imaging". In: *Med Image Anal*. Vol. 9791. International Society for Optics and Photonics. 2016, p. 979118.

[136] Jerome Faist et al. "Quantum cascade laser". In: *Science* 264.5158 (1994), pp. 553–556.

[137] Michael J. Pilling, Alex Henderson and Peter Gardner. *Quantum Cascade Laser Spectral Histopathology: Breast Cancer Diagnostics Using High Throughput Chemical Imaging*. Ed. by zenodo. 19th June 2017. URL: `https://zenodo.org/record/808456#.XBOM6mj7SUm` (visited on 14/12/2018).

[138] Charles M. Perou et al. "Molecular portraits of human breast tumours". In: *Nature* 406.6797 (2000), p. 747.

[139] Therese Sørlie et al. "Repeated observation of breast tumor subtypes in independent gene expression data sets". In: *PNAS* 100.14 (2003), pp. 8418–8423.

[140] Marlene Wolf et al. "Cathepsin D specifically cleaves the chemokines macrophage inflammatory protein-1$\alpha$, macrophage inflammatory protein-1$\beta$, and SLC that are expressed in human breast cancer". In: *Am. J. Pathol.* 162.4 (2003), pp. 1183–1190.

[141] Rokaya El Ansari et al. "The multifunctional solute carrier 3A2 (SLC3A2) confers a poor prognosis in the highly proliferative breast cancer subtypes". In: *Br. J. Cancer* (2018), p. 1.

[142] Catherine Kendall et al. "Vibrational spectroscopy: a clinical tool for cancer diagnostics". In: *Analyst* 134.6 (2009), pp. 1029–1045.

[143] Judith R. Mourant et al. "Biochemical differences in tumorigenic and nontumorigenic cells measured by Raman and infrared spectroscopy". In: *J Biomed Opt* 10.3 (2005), p. 031106.

[144] Agilent. *Cary 620 FTIR Chemical Imaging System*. 2018. URL: https://www.agilent.com/en-us/products/ftir/ftir-microscopes-imaging-systems/cary-620-ftir-microscopes/technology-benefits (visited on 06/11/2018).

[145] Agilent. *Optimized FTIR Spectrometer design for optical performance*. Tech. rep. Agilent Technologies, Inc, 2011. URL: https://www.agilent.com/cs/library/technicaloverviews/public/si-1335.pdf (visited on 06/11/2018).

[146] Agilent. *Cary 660 FTIR Spectrometer*. URL: https://www.agilent.com/en/products/ftir/ftir-benchtop-systems/cary-660-ftir-spectrometer (visited on 06/11/2018).

[147] John M. Chalmers and Neil J. Everall. "Qualitative and quantitative analysis of plastics, polymers and rubbers by vibrational spectroscopy". In: *Handbook of vibrational spectroscopy*. 2006.

# Experimental

The methods used to perform the plastic FTIR analysis are outlined. An introduction to the equipment, as well as some key notes of using the equipment are given.

**Agilent 620 microscope & Agilent 660 FTIR spectrometer**

FTIR images were collected using an Agilent 620 microscope connected to an Agilent 660 FTIR spectrometer, seen in Figure 12.1.

Cary 620 is a Focal Plane Array (FPA) chemical imaging FTIR microscope. It provides micron level spatially resolved, full FTIR spectra across a 2-D Field Of View (FOV), thus allowing high-definition chemical images to be created, for easy and detailed molecular chemical distribution visualisation. In high magnification mode, can measure with a 1.1 micron pixel size [144].

The Cary 660 FTIR is only available in conjunction with the Cary 610 or Cary 620 FTIR Microscope systems. This bench-top FTIR system comes with innovation optics design to improve the performance of spectrometers [145]. High spectral resolution, fast kinetics speeds and high signal-to-noise (S/N) performance can be achieved [146].

A self-developed, step by step user instruction of the machine is attached in the Appendix.

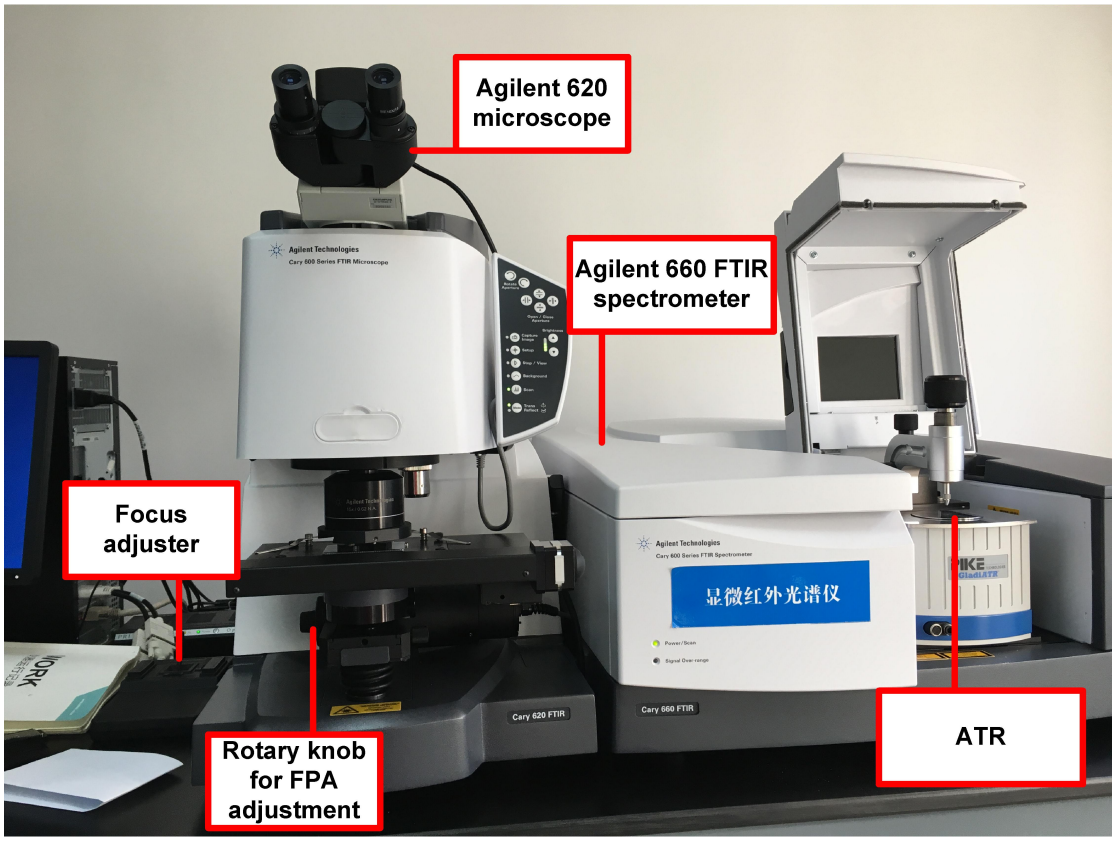



**Figure 12.1:** Agilent 620 microscope connected to an Agilent 660 FTIR spectrometer, with all the key component labelled. Detailed instruction of using 'focus adjuster' and 'rotary knob for FPA adjustment' are given in Appendix 1.

FTIR imaging in the region of interest can be performed using a mounted liquid nitrogen cooled 32 x 32 focal plane array (FPA) detector. Transmission in 'High magnification' mode can give pixel sizes of 1.1 $\mu m^2$, which together with the FPA detector providing a field of view of approximately 35 $\mu$m $\times$ 35 $\mu$m. Different combination of FPA mapping can be applied.

Using the 32 $\times$ 32 focal plane array detector, a long measurement time is needed. For a four by four image (128 $\times$ 128 pixels), using the setting of 1.1 $\mu m^2$ pixels

size, 2 $cm^{-1}$ spectral resolution, 16 sample scan, it takes up to 90 minutes to measure the 128 $\times$ 128 spectra.

**Presser**

For thermoplastics, hot compression moulding is a very convenient and effective means of preparing a free-standing film appropriate to an infrared transmission examination [147].

The IR absorption intensity is directly proportional to the sample thickness. Over thickness can cause baseline problems in FTIR spectra. For FTIR imaging analysis, the plastic composites need to be pressed into a thin film. Plastic samples obtained for FTIR study are plastic pellets, which are the result of a corotating twin-screw extruder. This section provides a detailed discussion of the pressing procedure, as well as the equipment used.

The presser (Figure 12.3) used can provide up to 250 $^{o}$C, 130 kg to the sample. Using the Atlas Constant Thickness and High Temp. Film Maker (Figure 12.2), P/N GS15640, the best PP sample thickness achieved is 24-25 $\mu$m.

A standard film making procedure is given:

1. Press the pellet into a flat plate using a hand presser;

2. Cut the flat circle into 2$\times$2 $mm^2$ square size to maintain a reasonable amount of plastic sample is used for film making;

3. The sample is put into the middle of the bottom foils before the top foils and upper assembly are added on top;

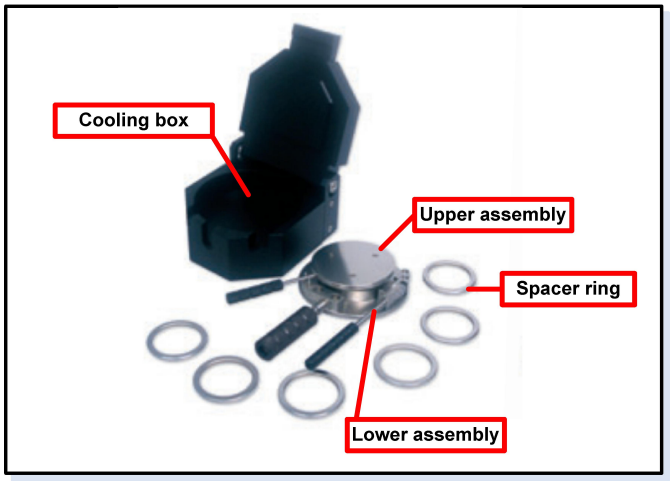



**Figure 12.2:** Specac P/N GS15640 High Temp. Film Maker, with all the key component labelled.

4. Place the whole assembly between the preheated platen surfaces for 10 minutes to melt the sample within the film maker accessory prior to compression via the press;

5. Press using three-stage auto-pressing setting;

6. Cool the whole assemble in the cooling box before the sample can be carefully peeled from the foils;

The detailed step by step instruction of the film making procedure is given in the Appendix.

Putting aluminium foil can improve the performance of the film making. Table 12.1 is the performance of the film making process with different parameter setting of the presser. Five points were selected to measure the result film's thickness. The range given in the thickness column indicates the lowest and the highest value of those measurements.

To maintain minimum influence, we aim to find the pressing setting that has

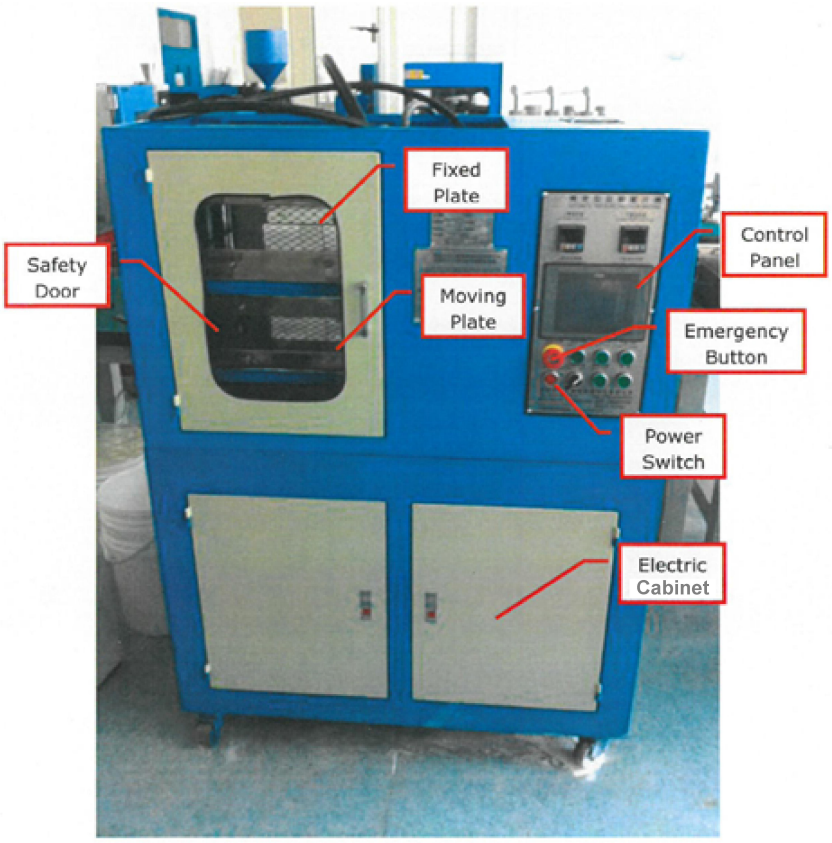



**Figure 12.3:** ZHENGGONG Equipment, ZG-100T presser, with all the key component labelled.

the lowest temperature, shortest pressing times that give the best thin, even film. The selected parameter combination is given in Table 12.1, experiment 3.

In the Table, experiment 1 produce an uneven film sample. Shorting the pressing time in step one (T1) and increasing the pre-heat time (experiment 2), the evenness of the sample is improved. Further extending the preheating and step one pressing time can produce a thinner film (experiment 3). Experiment 4-6 is the repeat of the best performance setting to check the repeatability. The results indicate that using the film making procedure, even thin plastic films can be produced.

| NO. | Sample | Presser configureation | | | | | | | | Thickness ($\mu$m) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Temperature ($^o$C) | Pre-heat | P1 | T1 | P2 | T2 | P3 | T3 | |
| 1 | PP | 220 | 120 | 15 | 600 | 30 | 120 | 40 | 120 | 15-37 |
| 2 | PP | 220 | 300 | 15 | 120 | 30 | 120 | 40 | 120 | 27-32 |
| 3 | PP | 220 | 600 | 15 | 300 | 30 | 120 | 40 | 120 | 24-25 |
| 4 | PP | 220 | 600 | 15 | 300 | 30 | 120 | 40 | 120 | 23-26 |
| 5 | PP | 220 | 600 | 15 | 300 | 30 | 120 | 40 | 120 | 24-27 |
| 6 | PP | 220 | 600 | 15 | 300 | 30 | 120 | 40 | 120 | 22-25 |

**Table 12.1:** Presser configuration. The temperature parameter is the temperature of the upper and lower pressing surface (ZG-100T presser). A period of re-heating is allowed to heat the pressing film maker assembly. The automatic mode of the presser is a three-stage pressing process. P1, T1, P2, T2, P3 and T3 are the pressure and time setting for each of the three steps. P1, P2 and P3 are in the unit of 'kg', while T1, T2 and T3 are using the unit of 'second'.

The fracture surfaces of the result PP film were scanned using a Zeiss Sigma scanning electron microscope (SEM) at an acceleration voltage of 20.00 kV. Before SEM analysis, the surfaces were coated with a thin layer of gold. The SEM image (Figure 12.4) shows the surface of the VPP film is smooth and flat.

One thing worth noting is the component percentage changing of the material will lead to the changing of mechanical properties, which will further lead to the different pressing requirements. For instance, pure VPP sample requires using 220 $^o$C, 40 kg pressing to get a 24-27 $\mu$m thickness film. Using the same setting, 70% VPP and 30% talc plastic sample gave the thickness 38-40 $\mu$m. The

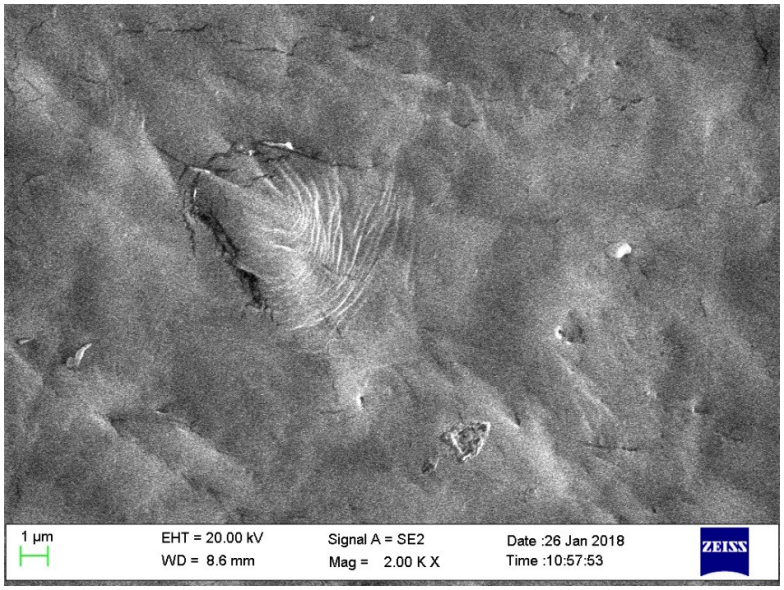



**Figure 12.4:** SEM image of the VPP film surface.

pressing parameters in Table 12.1 is for pure VPP sample only. The process of parameter configuration is needed for each new plastic sample.

# Appendix

## Appendix 1

Self-developed, step by step user instruction for Agilent 620 microscope connected to an Agilent 660 FTIR spectrometer:

- Add liquid nitrogen to the machine to make the detector temperature 78-79 K
- Create a new folder name as 'vpp-battery-scan16-Resolution4-1.1um', the name should not include any '.' in it
- Focus
  - Choose the colour button to start the process, a new dialogue windows will show up;
  - Set the parameters: scan number, spatial resolution, spectral resolution, WN range etc.
  - Choose 'Visible image' in the software;
  - Change the mode to reflection;
  - Change the focus adjuster to get a clear image of sample surface;

– Live FPA

- Change the mode back to trans;
- Remove the sample;
- Choose 'Live FPA' in the software;
- Click 'raw data' button;
- Change the rotary knob on the machine (underneath of the objective table) and the value setting bar to get a flat and intense line around 70%;
- Calibrate;
- Click 'OK' button in the software;

– Background scan

- Click the colour button again;
- Click 'background scan' in the software;
- Set the path for saving the file;
- Save the background file as 'background';
- Wait until the scan finished, the background abs value should be around 0.1;

– Sample scan

- Click 'Visible Image' in the dialogue windows;
- Enable $\rightarrow$ Set Conner 1 $\rightarrow$ Set Conner 2 $\rightarrow$ Capture;
- Click 'Captured image', select the image area n $\times$ n;

- Click 'scan',
- Wait until the scan finished;

– Save the data in spc file

- Export → export spc file

– Save the spc file into csv file

- Open 'spc.spc' file saved;
- Select all the spectrum and save as 'spc.csv';

## Appendix 2

Self-developed, step by step instruction for making thin plastic sample film:

– Press the pellet into flat circle using a hand presser;

- Flat circle could increase the contacting surface meanwhile reduce the time that needed to pre-heat the plastic.

– Cut the flat circle into $2 \times 2$ mm$^2$ square size;

- This is about the right among of plastic we need.

– Cut four circle aluminium foil in the diameter of 30 mm;

- The aluminium foil comes with the pressing ring is a bit too big. It difficult to avoid the aluminium film that size to stuck in between the ring and the upper lid. We reduce the diameter to 30 mm. It is possible to re-use the aluminium foils. However, drapes on the foil surface could cause drape on the plastic film surface. We recommend using four new

aluminium foils for each pressing, as new foil has better surface.

– Make the 'sandwich';

- Two layers of aluminium foil under, two layers of aluminium upper with the square plastic in between.

– Pre-heat;

- Set the temperature of the pressing machine to 220 $^{o}$C. Wait until the pressing surface reach that temperature. Put the pressing kit (with the 'sandwich in the kit') on the heated surface for 10 minutes.

– Auto-press;

- Set the parameters as following image (Figure 12.5) and start auto-pressing.
- Tips1: With the pressing kit in the machine, the door cannot be closed. The machine cannot start auto-press when the door is opened. Putting a metal bit (clamp) on the detector located on the bottom right conner of the door frame can solve this problem.
- Tips 2: Due to the pressure detector's sensitivity, before the pressing surface contacting each other, sometimes, the 'Current pressure' is having a value more than 10. With the 'Current pressure' value higher than 10, the auto-pressing couldn't be initialized. Manually rising or lowering the pressing surface can correct that value.

– Cooling;

- Normally a 5 minutes cooling in the cooling box is enough.

# Appendix 3

```
function [x,fval,exitflag,output] = fminsearch_cg(funfcn,x,options,varargin)
%FMINSEARCH Multidimensional unconstrained nonlinear minimization (Nelder-Mead).
%   X = FMINSEARCH(FUN,X0) starts at X0 and attempts to find a local minimizer
%   X of the function FUN.  FUN is a function handle.  FUN accepts input X and
%   returns a scalar function value F evaluated at X. X0 can be a scalar, vector
%   or matrix.
%
%   X = FMINSEARCH(FUN,X0,OPTIONS)  minimizes with the default optimization
%   parameters replaced by values in the structure OPTIONS, created
%   with the OPTIMSET function.  See OPTIMSET for details.  FMINSEARCH uses
%   these options: Display, TolX, TolFun, MaxFunEvals, MaxIter, FunValCheck,
%   PlotFcns, and OutputFcn.
%
%   X = FMINSEARCH(PROBLEM) finds the minimum for PROBLEM. PROBLEM is a
%   structure with the function FUN in PROBLEM.objective, the start point
%   in PROBLEM.x0, the options structure in PROBLEM.options, and solver
%   name 'fminsearch' in PROBLEM.solver.
%
%   [X,FVAL]= FMINSEARCH(...) returns the value of the objective function,
%   described in FUN, at X.
%
%   [X,FVAL,EXITFLAG] = FMINSEARCH(...) returns an EXITFLAG that describes
%   the exit condition. Possible values of EXITFLAG and the corresponding
%   exit conditions are
%
%    1  Maximum coordinate difference between current best point and other
%       points in simplex is less than or equal to TolX, and corresponding
%       difference in function values is less than or equal to TolFun.
%    0  Maximum number of function evaluations or iterations reached.
%   -1  Algorithm terminated by the output function.
%
%   [X,FVAL,EXITFLAG,OUTPUT] = FMINSEARCH(...) returns a structure
%   OUTPUT with the number of iterations taken in OUTPUT.iterations, the
%   number of function evaluations in OUTPUT.funcCount, the algorithm name
%   in OUTPUT.algorithm, and the exit message in OUTPUT.message.
%
%   Examples
%     FUN can be specified using @:
%        X = fminsearch(@sin,3)
%     finds a minimum of the SIN function near 3.
%     In this case, SIN is a function that returns a scalar function value
%     SIN evaluated at X.
%
%     FUN can be an anonymous function:
%        X = fminsearch(@(x) norm(x),[1;2;3])
%     returns a point near the minimizer [0;0;0].
%
%     FUN can be a parameterized function. Use an anonymous function to
%     capture the problem-dependent parameters:
%        f = @(x,c) x(1).^2+c.*x(2).^2;  % The parameterized function.
%        c = 1.5;                        % The parameter.
%        X = fminsearch(@(x) f(x,c),[0.3;1])
%
%   FMINSEARCH uses the Nelder-Mead simplex (direct search) method.
%
%   See also OPTIMSET, FMINBND, FUNCTION_HANDLE.

%   Reference: Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright,
%   Paul E. Wright, "Convergence Properties of the Nelder-Mead Simplex
%   Method in Low Dimensions", SIAM Journal of Optimization, 9(1):
```

```
61 %   p.112-147, 1998.
62
63 %   Copyright 1984-2017 The MathWorks, Inc.
64
65 % The following lines were changed by JK on 10/10/2018
66 % Get parameters of boundary conditions and more optimisation options
67 FminParams = varargin{1};
68 ExtraOptions = FminParams.ExtraOptions;
69 TolXI = ExtraOptions.TolXI;
70 cov = ExtraOptions.TolCOV;
71 params = FminParams;
72 % End of modifications
73
74 defaultopt = struct('Display','notify','MaxIter','200*numberOfVariables',...
75     'MaxFunEvals','200*numberOfVariables','TolX',1e-4,'TolFun',1e-4, ...
76     'FunValCheck','off','OutputFcn',[],'PlotFcns',[]);
77
78 % If just 'defaults' passed in, return the default options in X
79 if nargin==1 && nargout <= 1 && strcmpi(funfcn,'defaults')
80     x = defaultopt;
81     return
82 end
83
84 if nargin<3, options = []; end
85
86 % Detect problem structure input
87 if nargin == 1
88     if isa(funfcn,'struct')
89         [funfcn,x,options] = separateOptimStruct(funfcn);
90     else % Single input and non-structure
91         error('MATLAB:fminsearch:InputArg',...
92             getString(message('MATLAB:optimfun:fminsearch:InputArg')));
93     end
94 end
95
96 if nargin == 0
97     error('MATLAB:fminsearch:NotEnoughInputs',...
98         getString(message('MATLAB:optimfun:fminsearch:NotEnoughInputs')));
99 end
100
101
102 % Check for non-double inputs
103 if ~isa(x,'double')
104   error('MATLAB:fminsearch:NonDoubleInput',...
105     getString(message('MATLAB:optimfun:fminsearch:NonDoubleInput')));
106 end
107
108 n = numel(x);
109 numberOfVariables = n;
110
111 % Check that options is a struct
112 if ~isempty(options) && ~isa(options,'struct')
113     error('MATLAB:fminsearch:ArgNotStruct',...
114         getString(message('MATLAB:optimfun:commonMessages:ArgNotStruct', 3)));
115 end
116
117 printtype = optimget(options,'Display',defaultopt,'fast');
118 tolx = optimget(options,'TolX',defaultopt,'fast');
119 tolf = optimget(options,'TolFun',defaultopt,'fast');
120 maxfun = optimget(options,'MaxFunEvals',defaultopt,'fast');
```

```
121 maxiter = optimget(options,'MaxIter',defaultopt,'fast');
122 funValCheck = strcmp(optimget(options,'FunValCheck',defaultopt,'fast'),'on');
123
124 % In case the defaults were gathered from calling: optimset('fminsearch'):
125 if ischar(maxfun) || isstring(maxfun)
126     if strcmpi(maxfun,'200*numberofvariables')
127         maxfun = 200*numberOfVariables;
128     else
129         error('MATLAB:fminsearch:OptMaxFunEvalsNotInteger',...
130             getString(message('MATLAB:optimfun:fminsearch:↙
OptMaxFunEvalsNotInteger')));
131     end
132 end
133 if ischar(maxiter) || isstring(maxiter)
134     if strcmpi(maxiter,'200*numberofvariables')
135         maxiter = 200*numberOfVariables;
136     else
137         error('MATLAB:fminsearch:OptMaxIterNotInteger',...
138             getString(message('MATLAB:optimfun:fminsearch:↙
OptMaxIterNotInteger')));
139     end
140 end
141
142 switch printtype
143     case {'notify','notify-detailed'}
144         prnt = 1;
145     case {'none','off'}
146         prnt = 0;
147     case {'iter','iter-detailed'}
148         prnt = 3;
149     case {'final','final-detailed'}
150         prnt = 2;
151     case 'simplex'
152         prnt = 4;
153     otherwise
154         prnt = 1;
155 end
156 % Handle the output
157 outputfcn = optimget(options,'OutputFcn',defaultopt,'fast');
158 if isempty(outputfcn)
159     haveoutputfcn = false;
160 else
161     haveoutputfcn = true;
162     xOutputfcn = x; % Last x passed to outputfcn; has the input x's shape
163     % Parse OutputFcn which is needed to support cell array syntax for OutputFcn.
164     outputfcn = createCellArrayOfFunctions(outputfcn,'OutputFcn');
165 end
166
167 % Handle the plot
168 plotfcns = optimget(options,'PlotFcns',defaultopt,'fast');
169 if isempty(plotfcns)
170     haveplotfcn = false;
171 else
172     haveplotfcn = true;
173     xOutputfcn = x; % Last x passed to plotfcns; has the input x's shape
174     % Parse PlotFcns which is needed to support cell array syntax for PlotFcns.
175     plotfcns = createCellArrayOfFunctions(plotfcns,'PlotFcns');
176 end
177
178 header = ' Iteration   Func-count     min f(x)         Procedure';
```

```
179
180 % Convert to function handle as needed.
181 funfcn = fcnchk(funfcn,length(varargin));
182 % Add a wrapper function to check for Inf/NaN/complex values
183 if funValCheck
184     % Add a wrapper function, CHECKFUN, to check for NaN/complex values without
185     % having to change the calls that look like this:
186     % f = funfcn(x,varargin{:});
187     % x is the first argument to CHECKFUN, then the user's function,
188     % then the elements of varargin. To accomplish this we need to add the
189     % user's function to the beginning of varargin, and change funfcn to be
190     % CHECKFUN.
191     varargin = [{funfcn}, varargin];
192     funfcn = @checkfun;
193 end
194
195 n = numel(x);
196
197 % Initialize parameters
198 rho = 1; chi = 2; psi = 0.5; sigma = 0.5;
199 onesn = ones(1,n);
200 two2np1 = 2:n+1;
201 one2n = 1:n;
202
203 % Set up a simplex near the initial guess.
204 xin = x(:); % Force xin to be a column vector
205 v = zeros(n,n+1); fv = zeros(1,n+1);
206 v(:,1) = xin;    % Place input guess in the simplex! (credit L.Pfeffer at↙
Stanford)
207 x(:) = xin;    % Change x to the form expected by funfcn
208 fv(:,1) = funfcn(x,varargin{:});
209 func_evals = 1;
210 itercount = 0;
211 how = '';
212 % Initial simplex setup continues later
213
214 % Initialize the output and plot functions.
215 if haveoutputfcn || haveplotfcn
216     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'init',itercount, ...
217         func_evals, how, fv(:,1),varargin{:});
218     if stop
219         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
220         if  prnt > 0
221             disp(output.message)
222         end
223         return;
224     end
225 end
226 v0 = zeros(n,n+1);
227 v0(:,1)=xtransform(v(:,1),params);
228 % Print out initial f(x) as 0th iteration
229 if prnt == 3
230     disp(' ')
231     disp(header)
232     fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount, func_evals,↙
fv(1), how);
233 elseif prnt == 4
234     formatsave.format = get(0,'format');
235     formatsave.formatspacing = get(0,'formatspacing');
```

```
236     % reset format when done
237     oc1 = onCleanup(@()set(0,'format',formatsave.format));
238     oc2 = onCleanup(@()set(0,'formatspacing',formatsave.formatspacing));
239     format compact
240     format short e
241     disp(' ')
242     disp(how)
243     disp('v = ')
244     disp(v0)
245     disp('fv = ')
246     disp(fv)
247     disp('func_evals = ')
248     disp(func_evals)
249     disp('************************************************')
250
251     fileID=fopen('simplex.txt','w');
252     fprintf(fileID,'%6s\r\n','v = ');
253     for ii=1:size(v0,1)
254         fprintf(fileID,'%6g\t',v0(ii,:));
255         fprintf(fileID,'\r\n');
256     end
257       fprintf(fileID,'\r\n');
258     fprintf(fileID,'%6s\r\n','fv = ');
259     for ii=1:size(fv,1)
260         fprintf(fileID,'%6g\t',fv(ii,:));
261         fprintf(fileID,'\r\n');
262     end
263       fprintf(fileID,'\r\n');
264     fprintf(fileID,'%6s\r\n','func_evals = ');
265     fprintf(fileID,'%6g\t',func_evals);
266     fprintf(fileID,'\r\n');
267     fprintf(fileID,'%↙
6s\r\n','************************************************');
268 end
269 % OutputFcn and PlotFcns call
270 if haveoutputfcn || haveplotfcn
271     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'iter',itercount, ...
272         func_evals, how, fv(:,1),varargin{:});
273     if stop  % Stop per user request.
274         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
275         if  prnt > 0
276             disp(output.message)
277         end
278         return;
279     end
280 end
281
282 % Continue setting up the initial simplex.
283 % Following improvement suggested by L.Pfeffer at Stanford
284 usual_delta = 0.05;             % 5 percent deltas for non-zero terms
285 zero_term_delta = 0.00025;      % Even smaller delta for zero elements of x
286 for j = 1:n
287     y = xin;
288     if y(j) ~= 0
289         y(j) = (1 + usual_delta)*y(j);
290     else
291         y(j) = zero_term_delta;
292     end
293     v(:,j+1) = y;
```

```
294     x(:) = y; f = funfcn(x,varargin{:});
295     fv(1,j+1) = f;
296 end
297
298 % sort so v(1,:) has the lowest function value
299 [fv,j] = sort(fv);
300 v = v(:,j);
301 % transfer the v back to v0 for stopping criterion
302 v0 = zeros(n,n+1);
303 for j=1:n+1
304     v0(:,j)=xtransform(v(:,j),params);
305 end
306 how = 'initial simplex';
307 itercount = itercount + 1;
308 func_evals = n+1;
309 if prnt == 3
310     fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount, func_evals,↙
fv(1), how)
311 elseif prnt == 4
312     disp(' ')
313     disp(how)
314     disp('v = ')
315     disp(v0)
316     disp('fv = ')
317     disp(fv)
318     disp('func_evals = ')
319     disp(func_evals)
320     disp('TolXI = ')
321     disp(max(abs(v0(:,two2np1)-v0(:,onesn)),[],2))
322     disp('COV = ')
323     disp(std(fv)/mean(fv))
324     disp('****************************************************')
325
326     fprintf(fileID,'%6s\r\n','v = ');
327     for ii=1:size(v0,1)
328         fprintf(fileID,'%6.5g\t',v0(ii,:));
329         fprintf(fileID,'\r\n');
330     end
331       fprintf(fileID,'\r\n');
332     fprintf(fileID,'%6s\r\n','fv = ');
333     for ii=1:size(fv,1)
334         fprintf(fileID,'%6.5g\t',fv(ii,:));
335         fprintf(fileID,'\r\n');
336     end
337       fprintf(fileID,'\r\n');
338     fprintf(fileID,'%6s\r\n','func_evals = ');
339     fprintf(fileID,'%6g\t',func_evals);
340     fprintf(fileID,'\r\n');
341       fprintf(fileID,'\r\n');
342     fprintf(fileID,'%6s\r\n','TolXI = ');
343     tolxi_for_print=max(abs(v0(:,two2np1)-v0(:,onesn)),[],2);
344      for ii=1:size(tolxi_for_print,1)
345         fprintf(fileID,'%6.5g\t',tolxi_for_print(ii,:));
346         fprintf(fileID,'\r\n');
347      end
348     fprintf(fileID,'\r\n');
349     fprintf(fileID,'%6s\r\n','COV = ');
350     fprintf(fileID,'%6.5g\t',std(fv)/mean(fv));
351     fprintf(fileID,'\r\n');
352     fprintf(fileID,'%↙
```

```
6s\r\n','***************************************************');
353
354 end
355 % OutputFcn and PlotFcns call
356 if haveoutputfcn || haveplotfcn
357     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'iter',itercount, ...
358         func_evals, how, fv(:,1),varargin{:});
359     if stop  % Stop per user request.
360         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
361         if  prnt > 0
362             disp(output.message)
363         end
364         return;
365     end
366 end
367 exitflag = 1;
368
369 % Main algorithm: iterate until
370 % (a) the maximum coordinate difference between the current best point and the
371 % other points in the simplex is less than or equal to TolX. Specifically,
372 % until max(||v2-v1||,||v3-v1||,...,||v(n+1)-v1||) <= TolX,
373 % where ||.|| is the infinity-norm, and v1 holds the
374 % vertex with the current lowest value; AND
375 % (b) the corresponding difference in function values is less than or equal
376 % to TolFun. (Cannot use OR instead of AND.)
377 % The iteration stops if the maximum number of iterations or function evaluations
378 % are exceeded
379 while func_evals < maxfun && itercount < maxiter
380
381     COV=std(fv)/mean(fv);  %calculate the COV
382
383     simplex_size=max(abs(v0(:,two2np1)-v0(:,onesn)),[],2)-TolXI';  %distance↙
between the worst/next-to-worse point distance and best point, compare that with TolXI
384
385     if COV<cov
386         exitflag_reason=0;
387         break
388
389     elseif isempty(find(simplex_size>0))  %exit if the simplex size in all↙
dimension are samller than tolxi
390             exitflag_reason=1;
391         break
392     end
393
394     % Compute the reflection point
395
396     % xbar = average of the n (NOT n+1) best points
397     xbar = sum(v(:,one2n), 2)/n;
398     xr = (1 + rho)*xbar - rho*v(:,end);
399     x(:) = xr; fxr = funfcn(x,varargin{:});
400     func_evals = func_evals+1;
401
402     if fxr < fv(:,1)
403         % Calculate the expansion point
404         xe = (1 + rho*chi)*xbar - rho*chi*v(:,end);
405         x(:) = xe; fxe = funfcn(x,varargin{:});
406         func_evals = func_evals+1;
407         if fxe < fxr
408             v(:,end) = xe;
```

```
409             fv(:,end) = fxe;
410             how = 'expand';
411         else
412             v(:,end) = xr;
413             fv(:,end) = fxr;
414             how = 'reflect';
415         end
416     else % fv(:,1) <= fxr
417         if fxr < fv(:,n)
418             v(:,end) = xr;
419             fv(:,end) = fxr;
420             how = 'reflect';
421         else % fxr >= fv(:,n)
422             % Perform contraction
423             if fxr < fv(:,end)
424                 % Perform an outside contraction
425                 xc = (1 + psi*rho)*xbar - psi*rho*v(:,end);
426                 x(:) = xc; fxc = funfcn(x,varargin{:});
427                 func_evals = func_evals+1;
428 
429                 if fxc <= fxr
430                     v(:,end) = xc;
431                     fv(:,end) = fxc;
432                     how = 'contract outside';
433                 else
434                     % perform a shrink
435                     how = 'shrink';
436                 end
437             else
438                 % Perform an inside contraction
439                 xcc = (1-psi)*xbar + psi*v(:,end);
440                 x(:) = xcc; fxcc = funfcn(x,varargin{:});
441                 func_evals = func_evals+1;
442 
443                 if fxcc < fv(:,end)
444                     v(:,end) = xcc;
445                     fv(:,end) = fxcc;
446                     how = 'contract inside';
447                 else
448                     % perform a shrink
449                     how = 'shrink';
450                 end
451             end
452             if strcmp(how,'shrink')
453                 for j=two2np1
454                     v(:,j)=v(:,1)+sigma*(v(:,j) - v(:,1));
455                     x(:) = v(:,j); fv(:,j) = funfcn(x,varargin{:});
456                 end
457                 func_evals = func_evals + n;
458             end
459         end
460     end
461     % transfer the v back to v0 for stopping criterion
462     v0 = zeros(n,n+1);
463     for j=1:n+1
464         v0(:,j)=xtransform(v(:,j),params);
465     end
466 %     v0=
467 %     plot3([v0(1,1),v0(1,2)],[v0(2,1),v0(2,2)],[fv(1),fv(2)],'k','linewidth',1.5)
468 %     hold on
```

```
469 %     pause (1)
470 %     plot3([v0(1,2),v0(1,3)],[v0(2,2),v0(2,3)],[fv(2),fv(3)],'k','linewidth',1.5)
471 %     hold on
472 %     pause (1)
473 %     plot3([v0(1,1),v0(1,3)],[v0(2,1),v0(2,3)],[fv(1),fv(3)],'k','linewidth',1.5)
474 %     pause (1)
475     [fv,j] = sort(fv);
476     v = v(:,j);
477     v0 = v0(:,j);
478     itercount = itercount + 1;
479     if prnt == 3
480         fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount,↙
func_evals, fv(1), how)
481     elseif prnt == 4
482         disp(' ')
483         disp(how)
484         disp('v = ')
485         disp(v0)
486         disp('fv = ')
487         disp(fv)
488         disp('func_evals = ')
489         disp(func_evals)
490         disp('TolXI = ')
491         disp(max(abs(v0(:,two2np1)-v0(:,onesn)),[],2))
492         disp('COV = ')
493         disp(std(fv)/mean(fv))
494         disp('================================================')
495
496     fprintf(fileID,'%6s\r\n','v = ');
497     for ii=1:size(v0,1)
498         fprintf(fileID,'%6.5g\t',v0(ii,:));
499         fprintf(fileID,'\r\n');
500     end
501     fprintf(fileID,'\r\n');
502     fprintf(fileID,'%6s\r\n','fv = ');
503     for ii=1:size(fv,1)
504         fprintf(fileID,'%6.5g\t',fv(ii,:));
505         fprintf(fileID,'\r\n');
506     end
507     fprintf(fileID,'\r\n');
508     fprintf(fileID,'%6s\r\n','func_evals = ');
509     fprintf(fileID,'%6g\t',func_evals);
510     fprintf(fileID,'\r\n');
511       fprintf(fileID,'\r\n');
512     fprintf(fileID,'%6s\r\n','TolXI = ');
513     tolxi_for_print=max(abs(v0(:,two2np1)-v0(:,onesn)),[],2);
514      for ii=1:size(tolxi_for_print,1)
515         fprintf(fileID,'%6.5g\t',tolxi_for_print(ii,:));
516         fprintf(fileID,'\r\n');
517      end
518     fprintf(fileID,'\r\n');
519     fprintf(fileID,'%6s\r\n','COV = ');
520     fprintf(fileID,'%6.5g\t',std(fv)/mean(fv));
521     fprintf(fileID,'\r\n');
522     fprintf(fileID,'%↙
6s\r\n','================================================');
523
524     end
525     % OutputFcn and PlotFcns call
526     if haveoutputfcn || haveplotfcn
```

```
527         [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,↙
plotfcns,v(:,1),xOutputfcn,'iter',itercount, ...
528             func_evals, how, fv(:,1),varargin{:});
529         if stop  % Stop per user request.
530             [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
531             if  prnt > 0
532                 disp(output.message)
533             end
534             return;
535         end
536     end
537 end   % while
538 
539 x(:) = v(:,1);
540 fval = fv(:,1);
541 
542 output.iterations = itercount;
543 output.funcCount = func_evals;
544 output.algorithm = sprintf(['Nelder-Mead simplex direct search [MSimplex3\n',...
545     'InitFirstSimplexMethod = %d\nProcBoundaryMethod = %d\n',...
546     'ProcOutOfRangeMethod = %d\nExitCriteriaMethod = %d\n'],...
547     ExtraOptions.InitFirstSimplexMethod,ExtraOptions.ProcBoundaryMethod,...
548     ExtraOptions.ProcOutOfRangeMethod,ExtraOptions.ExitCriteriaMethod);
549 
550 % OutputFcn and PlotFcns call
551 if haveoutputfcn || haveplotfcn
552     callOutputAndPlotFcns(outputfcn,plotfcns,x,xOutputfcn,'done',itercount,↙
func_evals, how, fval, varargin{:});
553 end
554 
555 if func_evals >= maxfun
556     msg = getString(message('MATLAB:optimfun:fminsearch:ExitingMaxFunctionEvals',↙
sprintf('%f',fval)));
557     if prnt > 0
558         disp(' ')
559         disp(msg)
560     end
561     exitflag = 0;
562 elseif itercount >= maxiter
563     msg = getString(message('MATLAB:optimfun:fminsearch:ExitingMaxIterations',↙
sprintf('%f',fval)));
564     if prnt > 0
565         disp(' ')
566         disp(msg)
567     end
568     exitflag = 0;
569 else
570     if exitflag_reason==1
571         msg = ...
572             ['Optimization terminated: the current x satisfies the termination↙
criteria using OPTIONS.TolXI'];
573         if prnt > 1
574             disp(' ')
575             disp(msg)
576         end
577         exitflag = 1;
578     else
579         msg = ...
580             ['Optimization terminated: the current x satisfies the termination↙
criteria using OPTIONS.COV of ', sprintf('%e',cov)];
```

```
581         if prnt > 1
582             disp(' ')
583             disp(msg)
584         end
585         exitflag = 1;
586     end
587 end
588 
589 output.message = msg;
590 
591 %--------------------------------------------------------------------------
592 function [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,↙
plotfcns,x,xOutputfcn,state,iter,...
593     numf,how,f,varargin)
594 % CALLOUTPUTANDPLOTFCNS assigns values to the struct OptimValues and then calls↙
the
595 % outputfcn/plotfcns.
596 %
597 % state - can have the values 'init','iter', or 'done'.
598 
599 % For the 'done' state we do not check the value of 'stop' because the
600 % optimization is already done.
601 optimValues.iteration = iter;
602 optimValues.funccount = numf;
603 optimValues.fval = f;
604 optimValues.procedure = how;
605 
606 xOutputfcn(:) = x;  % Set x to have user expected size
607 stop = false;
608 state = char(state);
609 % Call output functions
610 if ~isempty(outputfcn)
611     switch state
612         case {'iter','init'}
613             stop = callAllOptimOutputFcns(outputfcn,xOutputfcn,optimValues,state,↙
varargin{:}) || stop;
614         case 'done'
615             callAllOptimOutputFcns(outputfcn,xOutputfcn,optimValues,state,varargin↙
{:});
616         otherwise
617             error('MATLAB:fminsearch:InvalidState',...
618                 getString(message('MATLAB:optimfun:fminsearch:InvalidState')));
619     end
620 end
621 % Call plot functions
622 if ~isempty(plotfcns)
623     switch state
624         case {'iter','init'}
625             stop = callAllOptimPlotFcns(plotfcns,xOutputfcn,optimValues,state,↙
varargin{:}) || stop;
626         case 'done'
627             callAllOptimPlotFcns(plotfcns,xOutputfcn,optimValues,state,varargin↙
{:});
628         otherwise
629             error('MATLAB:fminsearch:InvalidState',...
630                 getString(message('MATLAB:optimfun:fminsearch:InvalidState')));
631     end
632 end
633 
634 %--------------------------------------------------------------------------
```

```
635 function [x,FVAL,EXITFLAG,OUTPUT] = cleanUpInterrupt(xOutputfcn,optimValues)
636 % CLEANUPINTERRUPT updates or sets all the output arguments of FMINBND when the↙
optimization
637 % is interrupted.
638
639 % Call plot function driver to finalize the plot function figure window. If
640 % no plot functions have been specified or the plot function figure no
641 % longer exists, this call just returns.
642 callAllOptimPlotFcns('cleanuponstopsignal');
643
644 x = xOutputfcn;
645 FVAL = optimValues.fval;
646 EXITFLAG = -1;
647 OUTPUT.iterations = optimValues.iteration;
648 OUTPUT.funcCount = optimValues.funccount;
649 OUTPUT.algorithm = 'Nelder-Mead simplex direct search [MSimplex3]';
650 OUTPUT.message = getString(message('MATLAB:optimfun:fminsearch:↙
OptimizationTerminatedPrematurelyByUser'));
651
652 %--------------------------------------------------------------------------
653 function f = checkfun(x,userfcn,varargin)
654 % CHECKFUN checks for complex or NaN results from userfcn.
655
656 f = userfcn(x,varargin{:});
657 % Note: we do not check for Inf as FMINSEARCH handles it naturally.
658 if isnan(f)
659     error('MATLAB:fminsearch:checkfun:NaNFval',...
660         getString(message('MATLAB:optimfun:fminsearch:checkfun:NaNFval', localChar↙
( userfcn ))));
661 elseif ~isreal(f)
662     error('MATLAB:fminsearch:checkfun:ComplexFval',...
663         getString(message('MATLAB:optimfun:fminsearch:checkfun:ComplexFval',↙
localChar( userfcn ))));
664 end
665
666 %--------------------------------------------------------------------------
667 function strfcn = localChar(fcn)
668 % Convert the fcn to a character array for printing
669
670 if ischar(fcn)
671     strfcn = fcn;
672 elseif isstring(fcn) || isa(fcn,'inline')
673     strfcn = char(fcn);
674 elseif isa(fcn,'function_handle')
675     strfcn = func2str(fcn);
676 else
677     try
678         strfcn = char(fcn);
679     catch
680         strfcn = getString(message('MATLAB:optimfun:fminsearch:↙
NameNotPrintable'));
681     end
682 end
683 % ====================================
684 function xtrans = xtransform(x,params)
685 % converts unconstrained variables into their original domains
686
687 xtrans = zeros(params.xsize);
688 % k allows some variables to be fixed, thus dropped from the
689 % optimization.
```

```
690 k=1;
691 for i = 1:params.n
692   switch params.BoundClass(i)
693     case 1
694       % lower bound only
695       xtrans(i) = params.LB(i) + x(k).^2;
696
697       k=k+1;
698     case 2
699       % upper bound only
700       xtrans(i) = params.UB(i) - x(k).^2;
701
702       k=k+1;
703     case 3
704       % lower and upper bounds
705       xtrans(i) = (sin(x(k))+1)/2;
706       xtrans(i) = xtrans(i)*(params.UB(i) - params.LB(i)) + params.LB(i);
707       % just in case of any floating point problems
708       xtrans(i) = max(params.LB(i),min(params.UB(i),xtrans(i)));
709
710       k=k+1;
711     case 4
712       % fixed variable, bounds are equal, set it at either bound
713       xtrans(i) = params.LB(i);
714     case 0
715       % unconstrained variable.
716       xtrans(i) = x(k);
717
718       k=k+1;
719   end
720 end
721
```

# Appendix 4

```
function [x,fval,exitflag,output] = SMS3(funfcn,x,options,varargin)
% SMS3 Multidimensional unconstrained nonlinear minimisation using the Super↙
Modified Simplex Method


% The following lines were changed by JK on 10/10/2018
% Get parameters of boundary conditions and more optimisation options
FminParams = varargin{1};
ExtraOptions = FminParams.ExtraOptions;
TolXI = ExtraOptions.TolXI;
cov = ExtraOptions.TolCOV;
params = FminParams;
% End of modifications

defaultopt = struct('Display','notify','MaxIter','200*numberOfVariables',...
    'MaxFunEvals','200*numberOfVariables','TolX',1e-4,'TolFun',1e-4, ...
    'FunValCheck','off','OutputFcn',[],'PlotFcns',[]);

% If just 'defaults' passed in, return the default options in X
if nargin==1 && nargout <= 1 && strcmpi(funfcn,'defaults')
    x = defaultopt;
    return
end

if nargin<3, options = []; end

% Detect problem structure input
if nargin == 1
    if isa(funfcn,'struct')
        [funfcn,x,options] = separateOptimStruct(funfcn);
    else % Single input and non-structure
        error('MATLAB:fminsearch:InputArg',...
            getString(message('MATLAB:optimfun:fminsearch:InputArg')));
    end
end

if nargin == 0
    error('MATLAB:fminsearch:NotEnoughInputs',...
        getString(message('MATLAB:optimfun:fminsearch:NotEnoughInputs')));
end


% Check for non-double inputs
if ~isa(x,'double')
  error('MATLAB:fminsearch:NonDoubleInput',...
    getString(message('MATLAB:optimfun:fminsearch:NonDoubleInput')));
end

n = numel(x);
numberOfVariables = n;

% Check that options is a struct
if ~isempty(options) && ~isa(options,'struct')
    error('MATLAB:fminsearch:ArgNotStruct',...
        getString(message('MATLAB:optimfun:commonMessages:ArgNotStruct', 3)));
end

printtype = optimget(options,'Display',defaultopt,'fast');
tolx = optimget(options,'TolX',defaultopt,'fast');
tolf = optimget(options,'TolFun',defaultopt,'fast');
```

```
 60 maxfun = optimget(options,'MaxFunEvals',defaultopt,'fast');
 61 maxiter = optimget(options,'MaxIter',defaultopt,'fast');
 62 funValCheck = strcmp(optimget(options,'FunValCheck',defaultopt,'fast'),'on');
 63
 64 % In case the defaults were gathered from calling: optimset('fminsearch'):
 65 if ischar(maxfun) || isstring(maxfun)
 66     if strcmpi(maxfun,'200*numberofvariables')
 67         maxfun = 200*numberOfVariables;
 68     else
 69         error('MATLAB:fminsearch:OptMaxFunEvalsNotInteger',...
 70             getString(message('MATLAB:optimfun:fminsearch:↙
OptMaxFunEvalsNotInteger')));
 71     end
 72 end
 73 if ischar(maxiter) || isstring(maxiter)
 74     if strcmpi(maxiter,'200*numberofvariables')
 75         maxiter = 200*numberOfVariables;
 76     else
 77         error('MATLAB:fminsearch:OptMaxIterNotInteger',...
 78             getString(message('MATLAB:optimfun:fminsearch:↙
OptMaxIterNotInteger')));
 79     end
 80 end
 81
 82 switch printtype
 83     case {'notify','notify-detailed'}
 84         prnt = 1;
 85     case {'none','off'}
 86         prnt = 0;
 87     case {'iter','iter-detailed'}
 88         prnt = 3;
 89     case {'final','final-detailed'}
 90         prnt = 2;
 91     case 'simplex'
 92         prnt = 4;
 93     otherwise
 94         prnt = 1;
 95 end
 96 % Handle the output
 97 outputfcn = optimget(options,'OutputFcn',defaultopt,'fast');
 98 if isempty(outputfcn)
 99     haveoutputfcn = false;
100 else
101     haveoutputfcn = true;
102     xOutputfcn = x; % Last x passed to outputfcn; has the input x's shape
103     % Parse OutputFcn which is needed to support cell array syntax for OutputFcn.
104     outputfcn = createCellArrayOfFunctions(outputfcn,'OutputFcn');
105 end
106
107 % Handle the plot
108 plotfcns = optimget(options,'PlotFcns',defaultopt,'fast');
109 if isempty(plotfcns)
110     haveplotfcn = false;
111 else
112     haveplotfcn = true;
113     xOutputfcn = x; % Last x passed to plotfcns; has the input x's shape
114     % Parse PlotFcns which is needed to support cell array syntax for PlotFcns.
115     plotfcns = createCellArrayOfFunctions(plotfcns,'PlotFcns');
116 end
117
```

```
118 header = ' Iteration   Func-count     min f(x)         Procedure';
119
120 % Convert to function handle as needed.
121 funfcn = fcnchk(funfcn,length(varargin));
122 % Add a wrapper function to check for Inf/NaN/complex values
123 if funValCheck
124     % Add a wrapper function, CHECKFUN, to check for NaN/complex values without
125     % having to change the calls that look like this:
126     % f = funfcn(x,varargin{:});
127     % x is the first argument to CHECKFUN, then the user's function,
128     % then the elements of varargin. To accomplish this we need to add the
129     % user's function to the beginning of varargin, and change funfcn to be
130     % CHECKFUN.
131     varargin = [{funfcn}, varargin];
132     funfcn = @checkfun;
133 end
134
135 n = numel(x);
136
137 % Initialize parameters
138 rho = 1; chi = 2; psi = 0.5; sigma = 0.5;
139 onesn = ones(1,n);
140 two2np1 = 2:n+1;
141 one2n = 1:n;
142 safety_margin=0.3;
143 % Set up a simplex near the initial guess.
144 xin = x(:); % Force xin to be a column vector
145 v = zeros(n,n+1); fv = zeros(1,n+1);
146 v(:,1) = xin;    % Place input guess in the simplex! (credit L.Pfeffer at↙
Stanford)
147 x(:) = xin;    % Change x to the form expected by funfcn
148 fv(:,1) = funfcn(x,varargin{:});
149 func_evals = 1;
150 itercount = 0;
151 how = '';
152 % Initial simplex setup continues later
153
154 % Initialize the output and plot functions.
155 if haveoutputfcn || haveplotfcn
156     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'init',itercount, ...
157         func_evals, how, fv(:,1),varargin{:});
158     if stop
159         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
160         if  prnt > 0
161             disp(output.message)
162         end
163         return;
164     end
165 end
166 v0 = zeros(n,n+1);
167 v0(:,1)=xtransform(v(:,1),params);
168 % Print out initial f(x) as 0th iteration
169 if prnt == 3
170     disp(' ')
171     disp(header)
172     fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount, func_evals,↙
fv(1), how);
173 elseif prnt == 4
174     formatsave.format = get(0,'format');
```

```
175     formatsave.formatspacing = get(0,'formatspacing');
176     % reset format when done
177     oc1 = onCleanup(@()set(0,'format',formatsave.format));
178     oc2 = onCleanup(@()set(0,'formatspacing',formatsave.formatspacing));
179     format compact
180     format short e
181     disp(' ')
182     disp(how)
183     disp('v = ')
184     disp(v0)
185     disp('fv = ')
186     disp(fv)
187     disp('func_evals = ')
188     disp(func_evals)
189     disp('************************************************')
190 end
191 % OutputFcn and PlotFcns call
192 if haveoutputfcn || haveplotfcn
193     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'iter',itercount, ...
194         func_evals, how, fv(:,1),varargin{:});
195     if stop  % Stop per user request.
196         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
197         if  prnt > 0
198             disp(output.message)
199         end
200         return;
201     end
202 end
203
204 % Continue setting up the initial simplex.
205 % Following improvement suggested by L.Pfeffer at Stanford
206 usual_delta = 0.05;             % 5 percent deltas for non-zero terms
207 zero_term_delta = 0.00025;      % Even smaller delta for zero elements of x
208 for j = 1:n
209     y = xin;
210     if y(j) ~= 0
211         y(j) = (1 + usual_delta)*y(j);
212     else
213         y(j) = zero_term_delta;
214     end
215     v(:,j+1) = y;
216     x(:) = y; f = funfcn(x,varargin{:});
217     fv(1,j+1) = f;
218 end
219
220 % sort so v(1,:) has the lowest function value
221 [fv,j] = sort(fv);
222 v = v(:,j);
223 % transfer the v back to v0 for stopping criterion
224 v0 = zeros(n,n+1);
225 for j=1:n+1
226     v0(:,j)=xtransform(v(:,j),params);
227 end
228 how = 'initial simplex';
229 itercount = itercount + 1;
230 func_evals = n+1;
231 if prnt == 3
232     fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount, func_evals,↙
fv(1), how)
```

```
233 elseif prnt == 4
234     disp(' ')
235     disp(how)
236     disp('v = ')
237     disp(v0)
238     disp('fv = ')
239     disp(fv)
240     disp('func_evals = ')
241     disp(func_evals)
242     disp('TolXI = ')
243     disp(max(abs(v0(:,two2np1)-v0(:,onesn)),[],2))
244     disp('COV = ')
245     disp(std(fv)/mean(fv))
246     disp('***************************************************')
247 end
248 % OutputFcn and PlotFcns call
249 if haveoutputfcn || haveplotfcn
250     [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,plotfcns,v↙
(:,1),xOutputfcn,'iter',itercount, ...
251         func_evals, how, fv(:,1),varargin{:});
252     if stop  % Stop per user request.
253         [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
254         if  prnt > 0
255             disp(output.message)
256         end
257         return;
258     end
259 end
260 exitflag = 1;
261
262 % Main algorithm: iterate until
263 % (a) the coefficient of variation (COV), standard deviation of
264 % the responses in the simplex espressed as a percentage of the mean
265 while func_evals < maxfun && itercount < maxiter
266
267     COV=std(fv)/mean(fv);  %calculate the COV
268
269     simplex_size=max(abs(v0(:,two2np1)-v0(:,onesn)),[],2)-TolXI';  %distance↙
between the worst/next-to-worse point distance and best point, compare that with TolXI
270
271     if COV<cov
272         exitflag_reason=0;
273         break
274
275     elseif isempty(find(simplex_size>0))  %exit if the simplex size in all↙
dimension are samller than tolxi
276         exitflag_reason=1;
277         break
278     end
279
280     % Compute the reflection point
281
282     % xbar = average of the n (NOT n+1) best points
283     %make P bar
284     xbar = sum(v(:,one2n), 2)/n;
285     x(:) = xbar; fxbar = funfcn(x,varargin{:});
286     func_evals = func_evals+1;
287
288     p = xtransform(x,params);
289     fval_p=fxbar;
```

```
290
291     %make R
292     xr = (1 + rho)*xbar - rho*v(:,end);
293     x(:) = xr; fxr = funfcn(x,varargin{:});
294     func_evals = func_evals+1;
295
296     R = xtransform(x,params);
297     fval_r=fxr;
298
299     % evaluate Yopt and adjust for constraints
300     Yopt=(fv(:,end)-fxbar)/(fv(:,end)-2*fxbar+fxr)+0.5;
301     if Yopt>(1-safety_margin) && Yopt<=1
302         Yopt=1-safety_margin;
303     elseif Yopt<(1+safety_margin) && Yopt>1
304         Yopt=1+safety_margin;
305     elseif Yopt<safety_margin && Yopt>0
306         Yopt=safety_margin;
307     elseif Yopt>-safety_margin && Yopt<=0
308         Yopt=-safety_margin;
309     elseif Yopt>3
310         Yopt=3;
311     elseif Yopt<-1
312         Yopt=-1;
313     end
314
315     if (fv(:,end)-2*fxbar+fxr)<0
316         if fxr<=fv(:,end)
317         Yopt=3;
318         elseif fxr>fv(:,end)
319             Yopt=-1;
320         end
321     end
322     Yopt;
323     %make O
324     xO=Yopt*xbar+(1-Yopt)*v(:,end);
325     x(:) = xO; fxO = funfcn(x,varargin{:});
326     func_evals = func_evals+1;
327     x_opt=x;
328     fval_opt=fxO;
329
330     if fxO < fxr
331         % BNO
332         v(:,end) = xO;
333         fv(:,end) = fxO;
334         how = 'BNO'
335     else
336         v(:,end) = xr;
337         fv(:,end) = fxr;
338         how = 'BNR'
339     end
340     % transfer the v back to v0 for stopping criterion
341     v0 = zeros(n,n+1);
342     for j=1:n+1
343         v0(:,j)=xtransform(v(:,j),params);
344     end
345
346 %     plot3([v0(1,1),v0(1,2)],[v0(2,1),v0(2,2)],[fv(1),fv(2)],'k','linewidth',1.5)
347 %     hold on
348 %     pause (1)
349 %     plot3([v0(1,2),v0(1,3)],[v0(2,2),v0(2,3)],[fv(2),fv(3)],'k','linewidth',1.5)
```

```
350 %     hold on
351 %     pause (1)
352 %     plot3([v0(1,1),v0(1,3)],[v0(2,1),v0(2,3)],[fv(1),fv(3)],'k','linewidth',1.5)
353 %     pause (1)
354     [fv,j] = sort(fv);
355     v = v(:,j);
356     v0 = v0(:,j);
357     itercount = itercount + 1;
358     if prnt == 3
359         fprintf(' %5.0f        %5.0f     %12.6g         %s\n', itercount,↙
func_evals, fv(1), how)
360     elseif prnt == 4
361         disp(' ')
362         disp(how)
363         disp('v = ')
364         disp(v0)
365         disp('fv = ')
366         disp(fv)
367         disp('func_evals = ')
368         disp(func_evals)
369         disp('TolXI = ')
370         disp(max(abs(v0(:,two2np1)-v0(:,onesn)),[],2))
371         disp('COV = ')
372         disp(std(fv)/mean(fv))
373         disp('================================================')
374     end
375     % OutputFcn and PlotFcns call
376     if haveoutputfcn || haveplotfcn
377         [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,↙
plotfcns,v(:,1),xOutputfcn,'iter',itercount, ...
378             func_evals, how, fv(:,1),varargin{:});
379         if stop  % Stop per user request.
380             [x,fval,exitflag,output] = cleanUpInterrupt(xOutputfcn,optimValues);
381             if  prnt > 0
382                 disp(output.message)
383             end
384             return;
385         end
386     end
387 end   % while
388
389 x(:) = v(:,1);
390 fval = fv(:,1);
391
392 output.iterations = itercount;
393 output.funcCount = func_evals;
394 output.algorithm = sprintf(['Super modified simplex direct search↙
[SuperModSimplex3]\n',...
395     'InitFirstSimplexMethod = %d\nProcBoundaryMethod = %d\n',...
396     'ProcOutOfRangeMethod = %d\nExitCriteriaMethod = %d\n'],...
397     ExtraOptions.InitFirstSimplexMethod,ExtraOptions.ProcBoundaryMethod,...
398     ExtraOptions.ProcOutOfRangeMethod,ExtraOptions.ExitCriteriaMethod);
399
400 % OutputFcn and PlotFcns call
401 if haveoutputfcn || haveplotfcn
402     callOutputAndPlotFcns(outputfcn,plotfcns,x,xOutputfcn,'done',itercount,↙
func_evals, how, fval, varargin{:});
403 end
404
405 if func_evals >= maxfun
```

```
406     msg = getString(message('MATLAB:optimfun:fminsearch:ExitingMaxFunctionEvals',↙
sprintf('%f',fval)));
407     if prnt > 0
408         disp(' ')
409         disp(msg)
410     end
411     exitflag = 0;
412 elseif itercount >= maxiter
413     msg = getString(message('MATLAB:optimfun:fminsearch:ExitingMaxIterations',↙
sprintf('%f',fval)));
414     if prnt > 0
415         disp(' ')
416         disp(msg)
417     end
418     exitflag = 0;
419 else
420     if exitflag_reason==1
421         msg = ...
422             ['Optimization terminated: the current x satisfies the termination↙
criteria using OPTIONS.TolXI'];
423         if prnt > 1
424             disp(' ')
425             disp(msg)
426         end
427         exitflag = 1;
428     else
429         msg = ...
430             ['Optimization terminated: the current x satisfies the termination↙
criteria using OPTIONS.COV of ', sprintf('%e',cov)];
431         if prnt > 1
432             disp(' ')
433             disp(msg)
434         end
435         exitflag = 1;
436     end
437 end
438
439 output.message = msg;
440
441 %--------------------------------------------------------------------------
442 function [xOutputfcn, optimValues, stop] = callOutputAndPlotFcns(outputfcn,↙
plotfcns,x,xOutputfcn,state,iter,...
443     numf,how,f,varargin)
444 % CALLOUTPUTANDPLOTFCNS assigns values to the struct OptimValues and then calls↙
the
445 % outputfcn/plotfcns.
446 %
447 % state - can have the values 'init','iter', or 'done'.
448
449 % For the 'done' state we do not check the value of 'stop' because the
450 % optimization is already done.
451 optimValues.iteration = iter;
452 optimValues.funccount = numf;
453 optimValues.fval = f;
454 optimValues.procedure = how;
455
456 xOutputfcn(:) = x;  % Set x to have user expected size
457 stop = false;
458 state = char(state);
459 % Call output functions
```

```
460 if ~isempty(outputfcn)
461     switch state
462         case {'iter','init'}
463             stop = callAllOptimOutputFcns(outputfcn,xOutputfcn,optimValues,state,↙
varargin{:}) || stop;
464         case 'done'
465             callAllOptimOutputFcns(outputfcn,xOutputfcn,optimValues,state,varargin↙
{:});
466         otherwise
467             error('MATLAB:fminsearch:InvalidState',...
468                 getString(message('MATLAB:optimfun:fminsearch:InvalidState')));
469     end
470 end
471 % Call plot functions
472 if ~isempty(plotfcns)
473     switch state
474         case {'iter','init'}
475             stop = callAllOptimPlotFcns(plotfcns,xOutputfcn,optimValues,state,↙
varargin{:}) || stop;
476         case 'done'
477             callAllOptimPlotFcns(plotfcns,xOutputfcn,optimValues,state,varargin↙
{:});
478         otherwise
479             error('MATLAB:fminsearch:InvalidState',...
480                 getString(message('MATLAB:optimfun:fminsearch:InvalidState')));
481     end
482 end
483
484 %--------------------------------------------------------------------------
485 function [x,FVAL,EXITFLAG,OUTPUT] = cleanUpInterrupt(xOutputfcn,optimValues)
486 % CLEANUPINTERRUPT updates or sets all the output arguments of FMINBND when the↙
optimization
487 % is interrupted.
488
489 % Call plot function driver to finalize the plot function figure window. If
490 % no plot functions have been specified or the plot function figure no
491 % longer exists, this call just returns.
492 callAllOptimPlotFcns('cleanuponstopsignal');
493
494 x = xOutputfcn;
495 FVAL = optimValues.fval;
496 EXITFLAG = -1;
497 OUTPUT.iterations = optimValues.iteration;
498 OUTPUT.funcCount = optimValues.funccount;
499 OUTPUT.algorithm = 'Super modified simplex direct search [SuperModSimplex3]';
500 OUTPUT.message = getString(message('MATLAB:optimfun:fminsearch:↙
OptimizationTerminatedPrematurelyByUser'));
501
502 %--------------------------------------------------------------------------
503 function f = checkfun(x,userfcn,varargin)
504 % CHECKFUN checks for complex or NaN results from userfcn.
505
506 f = userfcn(x,varargin{:});
507 % Note: we do not check for Inf as FMINSEARCH handles it naturally.
508 if isnan(f)
509     error('MATLAB:fminsearch:checkfun:NaNFval',...
510         getString(message('MATLAB:optimfun:fminsearch:checkfun:NaNFval', localChar↙
( userfcn ))));
511 elseif ~isreal(f)
512     error('MATLAB:fminsearch:checkfun:ComplexFval',...
```

```
513         getString(message('MATLAB:optimfun:fminsearch:checkfun:ComplexFval',↙
localChar( userfcn ))));
514 end
515
516 %-----------------------------------------------------------------------
517 function strfcn = localChar(fcn)
518 % Convert the fcn to a character array for printing
519
520 if ischar(fcn)
521     strfcn = fcn;
522 elseif isstring(fcn) || isa(fcn,'inline')
523     strfcn = char(fcn);
524 elseif isa(fcn,'function_handle')
525     strfcn = func2str(fcn);
526 else
527     try
528         strfcn = char(fcn);
529     catch
530         strfcn = getString(message('MATLAB:optimfun:fminsearch:↙
NameNotPrintable'));
531     end
532 end
533
534 % ====================================
535 function xtrans = xtransform(x,params)
536 % converts unconstrained variables into their original domains
537
538 xtrans = zeros(params.xsize);
539 % k allows some variables to be fixed, thus dropped from the
540 % optimization.
541 k=1;
542 for i = 1:params.n
543   switch params.BoundClass(i)
544     case 1
545       % lower bound only
546       xtrans(i) = params.LB(i) + x(k).^2;
547
548       k=k+1;
549     case 2
550       % upper bound only
551       xtrans(i) = params.UB(i) - x(k).^2;
552
553       k=k+1;
554     case 3
555       % lower and upper bounds
556       xtrans(i) = (sin(x(k))+1)/2;
557       xtrans(i) = xtrans(i)*(params.UB(i) - params.LB(i)) + params.LB(i);
558       % just in case of any floating point problems
559       xtrans(i) = max(params.LB(i),min(params.UB(i),xtrans(i)));
560
561       k=k+1;
562     case 4
563       % fixed variable, bounds are equal, set it at either bound
564       xtrans(i) = params.LB(i);
565     case 0
566       % unconstrained variable.
567       xtrans(i) = x(k);
568
569       k=k+1;
570   end
```

```
571 end
572
573
```


**Figure 12.5:** Presser parameter setting